

Within-individual variation of selected clinical measurements

by

Alexander Gerard Gough MA VetMB

A thesis submitted to the University of Birmingham for the
degree of DOCTOR OF PHILOSOPHY

Institute of Applied Health Research

College of Medical and Dental Sciences

University of Birmingham

July 2024

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Background

Within-individual measured variation of a clinical test is the variation over time of an individual's clinical test results. Total measured within-individual variation includes biological variation and pre-analytic and analytic variation. Biological variation includes systematic variation, for example circadian or seasonal, and chance fluctuation around a homeostatic set point. Pre-analytic variation includes environmental and patient pre-measurement factors that affect the test result and analytic variation is the variation caused by the inherent accuracy of the test.

Within-individual variation is important in clinical decision-making because clinical decisions are informed by clinical test results and the existence of within-individual variation means that a single, or even multiple clinical measurements may not truly reflect the average status of the patient, which can lead to incorrect diagnostic, prognostic and therapeutic decisions.

Aims

This thesis aims to explore the existing data on within-individual variation on laboratory tests, functional measures and questionnaire-based tests and use large primary care databases to estimate this variation in these selected tests.

Methods

Four systematic reviews of the existing literature on within-individual variation of seven tests were performed according to PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines. Three tests were laboratory-based: HbA1c (Haemoglobin A1c), CRP (C-Reactive Protein) and hsCRP (high-sensitivity C-Reactive Protein); two were functional tests, the spirometry measures FEV1 (Forced Expiratory Volume in one second) and FVC (Forced vital capacity); three were questionnaire-based, the depression symptom measurement instruments PHQ-9 (Patient Health Questionnaire-9), BDI (Becks Depression Inventory) and HADS (Hospital Anxiety and Depression Scale). Four large retrospective cohort studies were then conducted using the large primary care databases, The IQVIA Medical Research Database (IMRD) and The Clinical Practice

Research Datalink (CPRD), to estimate the real-world within-individual measured variation of seven tests, HbA1c, CRP, FEV1, FVC, PHQ-9, HADS and BDI.

Results

For the four systematic reviews, in total, 8,724 abstracts were screened, 873 full texts were examined and 316 studies were included.

For the HbA1c systematic review, median within-individual coefficient of variation (CV) of HbA1c was 0.070 (IQR 0.034 to 0.09). For the CRP/hsCRP systematic review, median CV for CRP was 0.41 (range 0.11 to 0.89), and for hsCRP, median CV was 0.44 (range 0.27 to 0.76). For the spirometry systematic review, the median CV of FEV1 was 0.034 (IQR 0.025 to 0.067) and of FVC was 0.031 (0.02 to 0.048). For the depression symptom measurement instruments review, the pooled estimates of ICCs (Intraclass Correlations) and 95% confidence intervals for BDI, HADS, PHQ-9 and PHQ-2 were 0.89 (0.84-0.93), 0.89 (0.85-0.93), 0.84 (0.81-0.88) and 0.75 (0.56-0.93), respectively.

For the HbA1c retrospective cohort study, the overall measured within-individual coefficient of variation for HbA1c (CVT) was 0.20 (95%CI 0.20 to 0.20). For the CRP retrospective cohort study, the overall measured coefficient of variation for CRP was 1.148 (95% CI 1.147 to 1.149). For the spirometry retrospective cohort study, the overall coefficient of variation for FEV1 was 0.224 (95% CI 0.221 to 0.228) and for FVC was 0.152 (0.150 to 0.155). For the depression symptom measurement instrument retrospective cohort study, the overall measured coefficient of variation (95% CI) was 0.412 (0.411 to 0.413) for PHQ-9, 0.370 (0.365 to 0.374) for HADS and 0.763 (0.701 to 0.826) for BDI. Measured variation tended to increase with worsening disease severity except for the depression symptom measurement instruments.

Discussion

The systematic reviews in this thesis showed that there was a lack of good quality, generalisable data on the within-individual measured variation of a variety of clinical measurements, and this was particularly pronounced for non-laboratory tests. The retrospective cohort studies showed that the estimated measured within-individual variation for the selected tests was much higher than previously reported. They also showed that with the exception of the depression symptom measurement instruments, variation increased with worsening disease severity.

Limitations of the systematic reviews in this thesis include difficulties in identifying data on variation in studies which were not primarily designed to report variation. Limitations with the cohort studies include the difficulty of accounting for confounding by indication, and inaccuracies in the database.

Conclusion

This thesis demonstrates that the current published amount and quality of data on within-individual measured variation in real-world settings is limited, especially for non-laboratory tests. The estimates of within-individual measured variation from the analyses of primary care records in this thesis are higher than published estimates of within-individual variability. This may have important consequences for the accuracy of clinical decision-making with respect to diagnosis, prognosis and management. The extent of measured within-individual variation should be considered by clinicians, clinical trial authors and the designers of drug trials. Further research is needed on the extent of measured within-individual variation in a wide range of tests in real-world settings and its effect on clinical decision-making. Research into methods to account for confounding by indication would be helpful.

Dedication

To all my family for support throughout this process, especially Naomi and Abigail and my mother Joan Gough.

But above all to my father, Ian Gough (1940 – 2021) for all his encouragement throughout my education and career.

Acknowledgements

First and foremost, for excellent support, guidance and mentoring throughout this long process, thank you to my supervisors Professor Tom Marshall and Dr Alice Sitch. Thanks also to my sister Dr Amanda Gough, whose insight as a practising GP has been invaluable. Thanks to other current and former members of the Institute of Applied Health Research in the Medical and Dental School of Birmingham University for help with various aspects of this project including: Simon Baldwin and Erica Ferris for donating a large amount of time in screening systematic review papers and ensuring data accuracy; Dr Krishna Ghokale, Dr Anuradhaa Subramanian, Dr Siang Ing Lee and Dr Naijie Guan for assistance with the use of Dexter to extract data from the THIN and CPRD databases; and Dr Sophie Beese for reviewing the systematic review search strategies; Konstantinos Tryposkiadis for assistance with the statistical methods of the systematic reviews; Professor Alice Turner for advice and comments on the spirometry chapters and Dr Shamil Haroon for general advice on spirometry; Dr Funke Akiboye for assistance with queries about HbA1c; my fellow students for support and assistance during the pandemic, including but not limited to Dr Ciara Harris, Dr Ceren Çağlıyangil, Dr Sonja Bloch, Dr Clare MacDonald, Dr Katherine Woolley, Dr An Te and Dr Tolulope Olufunlayo.

1 Contents

1	Chapter one: Introduction	32
1.1	Clinical measurement and within-individual measurement variability.....	32
1.1.1	What is measured in clinical settings?	33
1.1.2	What are clinical measurements used for?	36
1.2	What are the sources of variability in a measurement?	37
1.2.1	Biological variation	39
1.2.2	Pathological variation.....	39
1.2.3	Pre-analytical variability	41
1.2.4	Analytical variation	42
1.2.5	Post-analytical variation (Interpretation and clinical decision-making)	45
1.2.6	Total within-individual variation	45
1.3	Terminology	46
1.4	Why is within-individual variation a problem?.....	49
1.4.1	Example 1:Blood pressure	49
1.4.2	Example 2: C-reactive protein.....	50
1.4.3	Example 3: Estimated glomerular filtration rate and albumin creatinine ratio	51
1.5	Testing in primary care	51
1.5.1	How often are tests repeated in general practice?.....	52
1.6	The effect of within-individual measurement variation on test results in practice	57

1.6.1	The problems with reference ranges	58
1.6.2	Repeat testing.....	59
1.7	The importance of within-individual measurement variation on decision-making in general practice	60
1.7.1	Existing variation data.....	62
1.8	Sources of biological variability	63
1.8.1	Variability over lifespan	64
1.8.2	Circadian or diurnal rhythms	64
1.8.3	Monthly cycles	65
1.8.4	Seasonal cycles	65
1.8.5	Non-systematic or chance variability	65
1.9	Consistency of measures of variation	66
1.10	Measures of variation and their definitions.....	68
1.11	Research question, aims and hypotheses	69
1.12	Methodology overview	71
1.13	Structure of thesis	72
2	Chapter two: Within-individual variation of HbA1c: A systematic review and meta-analysis	74
2.1	Abstract.....	74
2.2	Introduction	76
2.3	Aim of review	78
2.4	Methods	79

2.5	Results.....	83
2.6	Discussion.....	94
2.7	Conclusion	100
3	Chapter three: Within-individual variation of C-reactive protein and high-sensitivity C-reactive protein: A systematic review and meta-analysis	101
3.1	Abstract.....	101
3.2	Introduction	103
3.3	Aim of review	106
3.4	Methods	106
3.5	Results.....	109
3.6	Discussion.....	117
3.7	Conclusion	121
4	Chapter four: Within-individual variation in respiratory function tests - A Systematic Review and Meta-analysis	123
4.1	Abstract.....	123
4.2	Aim of review	125
4.3	Introduction	125
4.4	Methods	128
4.5	Results.....	131
4.6	Discussion.....	147
4.7	Conclusion	151

5	Chapter five: Within-individual Variation of Measured Depression Symptoms: A systematic review and meta-analysis	153
5.1	Abstract.....	153
5.2	Aim of review	155
5.3	Introduction	155
5.4	Methods	157
5.5	Results.....	160
5.6	Discussion.....	170
5.7	Conclusion	174
6	Chapter six: Within-individual Variation of HbA1c measurements in primary care: a retrospective cohort study	176
6.1	Abstract.....	176
6.2	Introduction	177
6.3	Aims.....	180
6.4	Methods	180
6.4.1	Study design and data sources.....	180
6.4.2	Ethical approval	181
6.4.3	Setting.....	182
6.4.4	Participants	182
6.4.5	Variables.....	182
6.4.6	Bias.....	183
6.4.7	Study size	184

6.4.8	Quantitative variables	184
6.4.9	Statistical methods.....	185
6.5	Results.....	187
	Participants.....	187
6.5.1	Descriptive data and outcome data	188
6.6	Discussion.....	194
6.6.1	Key results	194
6.6.2	Strengths and limitations.....	196
6.6.3	Interpretation and generalisability	198
6.7	Conclusions	201
7	Chapter seven: Within-individual Variation of CRP measurements in primary care: a retrospective cohort study	202
7.1	Abstract.....	202
7.2	Introduction	203
7.3	Aims.....	205
7.4	Methods	206
7.4.1	Participants	206
7.4.2	Variables.....	206
7.4.3	Quantitative variables	Error! Bookmark not defined.
7.4.4	Statistical methods.....	209
7.5	Results.....	210
	Participants.....	210

7.5.1	Descriptive data and outcome data	211
7.6	Discussion.....	216
7.6.1	Key results	216
7.6.2	Strengths and limitations.....	218
7.6.3	Interpretation and generalisability	220
7.7	Conclusions	223
8	Chapter eight: Within-individual variation of spirometry measurements in primary care: a retrospective cohort study	224
8.1	Abstract.....	224
8.2	Introduction	225
8.3	Aims.....	228
8.4	Methods	229
8.4.1	Participants.....	229
8.4.2	Variables.....	229
8.4.3	Bias and study size	230
8.4.4	Quantitative variables	Error! Bookmark not defined.
8.4.5	Statistical methods.....	231
8.5	Results.....	234
	Participants.....	234
8.5.1	Descriptive data and outcome data	235
8.6	Discussion.....	244
8.6.1	Key results	244

8.6.2	Strengths and limitations.....	245
8.6.3	Interpretation and generalisability	247
8.7	Conclusions	250
9	Chapter nine: Within-individual variation of depression symptom measurement instruments in primary care: a retrospective cohort study	251
9.1	Abstract.....	251
9.2	Introduction	253
9.3	Aims.....	256
9.4	Methods	257
9.4.1	Study design	257
9.4.2	Data sources.....	257
9.4.3	Ethical approval	258
9.4.4	Setting.....	258
9.4.5	Participants	258
9.4.6	Variables.....	259
9.4.7	Bias and Study size	260
9.4.8	Quantitative variables	Error! Bookmark not defined.
9.4.9	Statistical methods.....	260
9.5	Results.....	262
	Participants.....	262
9.5.1	Descriptive data and outcome data	263
9.6	Discussion.....	271

9.6.1	Key results	271
9.6.2	Strengths and limitations.....	273
9.6.3	Interpretation and generalisability	275
9.7	Conclusions	278
10	Chapter ten: Discussion	280
10.1	Main findings: summary and implications for practice	280
10.1.1	Overall summary.....	280
10.1.2	Systematic reviews	282
10.1.3	IMRD and CPRD database studies	291
10.2	Comparison to other research.....	304
10.2.1	Systematic reviews in this thesis	304
10.2.2	Analogous studies	304
10.3	Applications of Research	305
10.3.1	Current initiatives	306
10.4	Contribution of the research findings	308
10.5	Research Recommendations.....	321
10.6	Conclusion	325
11	Declarations.....	327
12	Appendix 2 HbA1c systematic review appendices	328
12.1	Appendix 2.1 Search strategies	328
12.2	Appendix 2.2 Databases searched	333

12.3	Appendix 2.3a Risk of bias template and explanatory notes.....	334
12.4	Appendix 2.3b Alternative risk of bias template	337
12.5	Appendix 2.4 Included studies and characteristics	340
12.6	Appendix 2.5 Risk of bias estimates for each paper	345
12.7	Appendix 2.6 Statistical calculations for ICC meta-analysis.....	348
12.7.1	Stata code for ICC:	348
12.7.2	Stata code for CV	348
12.8	Appendix 2.7 Forest plots of subgroup and sensitivity analyses	351
13	Appendix 3 CRP systematic review appendices.....	355
13.1	Appendix 3.1 Statistical calculations for ICC meta-analysis.....	355
13.2	Appendix 3.2 Study characteristics	358
13.3	Appendix 3.3 Risk of bias scores	362
13.4	Appendix 3.4 Forest plots of subgroup analyses	364
13.4.1	CRP CVs	364
13.4.2	hsCRP CVs.....	368
13.4.3	CRP ICCs	371
13.4.4	hsCRP ICCs	372
14	Appendix 4 Spirometry systematic review appendices.....	373
14.1	Appendix 4.1 Search strategies	373
14.2	Appendix 4.2 Variables extracted from studies	377
14.3	Appendix 4.3 Forest plots of subgroup analyses	378

15	Appendix 5 Depression symptom measurement instrument systematic reviews	388
15.1	Description of Depression symptom measurement instruments	388
15.2	Appendix 5.2 Databases searched	391
15.3	Appendix 5.3 Search strategies	392
15.4	Appendix 5.4 Included studies and characteristics	395
15.5	Appendix 5.5 Risk of bias scores for all studies	399
15.6	Appendix 5.6 Heterogeneity results	400
16	Appendix 6 HbA1c cohort study appendices	401
16.1	Appendix 6.1 - Subgroup analysis	401
16.2	Appendix 6.2 – sensitivity analyses	409
17	Appendix 7 CRP cohort study appendices	417
17.1	Appendix 7.1 - Subgroup analysis	417
17.2	Appendix 7.2 – sensitivity analyses	424
18	Appendix 8 Spirometry cohort study appendices.....	432
18.1	Appendix 8.1 - Subgroup analysis	432
18.2	Appendix 8.2 – sensitivity analyses	443
19	Appendix 9 Depression measurement instrument appendices.....	459
19.1	Appendix 9.1 - Subgroup analysis	459
19.2	Appendix 9.2 – sensitivity analyses	464
20	References	497

FIGURES

Figure 1.1 Categories of clinical measurements with examples	35
Figure 1.2 Sources of within-individual variability in clinical measurements and decision-making	38
Figure 1.3 Illustration of bias, precision and accuracy, using average distance to target centre of five shots	43
Figure 1.4 Terminology used to describe variation.....	48
Figure 1.5 Frequency of selected tests performed in a sample of GP practices taken from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.	53
Figure 1.6 Median number of repeats in the same individual of a selection of tests performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.	54
Figure 1.7 Distribution of number of repeated tests of HbA1c in the same individual performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.	55
Figure 1.8 Distribution of number of repeated tests of forced vital capacity (FVC) in the same individual performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.	56
Figure 1.9 Distribution of number of repeated tests of C-reactive protein (CRP) in the same individual performed by a group of practices from the IMRD database. Data	

from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.....	57
Figure 2.1 PRISMA flow diagram describing selection of studies.	83
Figure 2.2 Forest plot of CV for studies reporting in units of %, split by health status.	89
Figure 2.3 Forest plot of CV for studies reporting in units of mmol/mol, split by health status.....	90
Figure 2.4 Forest plot of SD for studies reporting in units of %, split by health status.	91
Figure 2.5 Forest plot with meta-analysis of ICC for studies reporting units of %, split by health status.	92
Figure 2.6 Scatter plot of mean population HbA1c against CV, with line of best fit. .	93
Figure 3.1 PRISMA flow diagram describing selection of studies	109
Figure 3.2 Modified forest plot of all studies that reported a CV for CRP, split by health status.....	113
Figure 3.3 Modified forest plot of all studies that reported a CV for hsCRP, split by health status.....	114
Figure 3.4 Forest plot of all studies that reported an ICC for CRP, split by health status, with meta-analysis.	115
Figure 3.5 Forest plot of all studies that reported an ICC for hsCRP, split by health status, with meta-analysis	116
Figure 4.1 PRISMA flow diagram describing selection of studies.	131
Figure 4.2 Forest plot of within-individual CV of FEV1 by health status, all studies using units of Litres.	143

Figure 4.3 Forest plot of within-individual CV of FVC by health status, all studies using units of Litres.	144
Figure 4.4 Forest plot and meta-analysis of ICCs of FEV1 by health status, all studies using units of Litres.	146
Figure 4.5 Forest plot and meta-analysis of ICCs of FVC by health status	147
Figure 6.1 Participant selection algorithm	187
Figure 6.2 Heat map showing the coefficient of variation expressed as % for multiple subgroup analysis according to sex, age, ethnicity, deprivation and BMI.	191
Figure 6.3 Coefficient of variation by patient median with 95% confidence intervals	191
Figure 6.4 Probability of second result being 5 mol/mol or more different than first result.	192
Figure 7.1 Participant selection algorithm	210
Figure 7.2 Heat map showing the coefficient of variation for multivariable analysis according to median CRP result, sex, age, ethnicity, deprivation and BMI.	213
Figure 7.3 Coefficient of variation by patient median with 95% confidence intervals	214
Figure 7.4 Probability of second result being 10mg/L different than first result.	215
Figure 8.1 Participant selection algorithm	234
Figure 8.2 Graph of CVT of FEV1 and FVC by respiratory status with 95% confidence intervals.....	238
Figure 8.3 Heat map showing the coefficient of variation of FEV1 expressed as a decimal fraction for multivariable analysis according to sex, age, smoking status and BMI.....	240

Figure 8.4 Heat map showing the coefficient of variation of FVC expressed as a decimal fraction for multivariable analysis according to sex, age, smoking status and BMI.....	240
Figure 8.5 Coefficient of variation of FEV1 by patient median with 95% confidence intervals.....	241
Figure 8.6 Coefficient of variation of FVC by patient median with 95% confidence intervals.....	241
Figure 8.7 Probability of second FEV1 result being 0.5l or more different than first result.	242
Figure 8.8 Probability of second FVC result being 0.5l or more different than first result.	242
Figure 9.1 Participant selection algorithm for PHQ-9	262
Figure 9.2 Overall CV of PHQ-9, HADS and BDI with 95% confidence intervals. ...	263
Figure 9.3 Coefficient of variation of PHQ-9 by patient median with 95% confidence intervals.....	264
Figure 9.4 Coefficient of variation of PHQ-9 score by depression diagnosis.....	267
Figure 9.5 Coefficient of variation of PHQ-9 by anti-depressant use.....	267
Figure 9.6 Heat map showing the coefficient of variation of PHQ-9 expressed as a decimal fraction for multivariable analysis according to age, ethnicity, comorbidities, depression diagnosis and PHQ-9 score. Age is in years.....	268
Figure 9.7 Probability of second PHQ-9 result based on first result.	269
Figure 9.8 Probability of second PHQ-9 score being 6 points or more greater or less than the first score(see methods).	270
Figure 10.1 Simplified decision tree for diagnosis and management of diabetes based on NICE guidelines[339].	311

Figure 10.2 Simulation of distribution of 1000 patients with a true mean HbA1c of 6.5% (48mmol/mol) and a CV of 0.20.	312
Figure 12.1 Forest plot of CV for studies reporting units of %, split by risk of bias .	351
Figure 12.2 Forest plot of CV for studies reporting units of %, split by setting.	352
Figure 12.3 Forest plot of CV for studies reporting units of %, split by short (S) or long (L) term measurements	353
Figure 12.4 Forest plot of CV for studies reporting units of %, split by whether CVI (I) or CVT (T) was reported or whether this was uncertain (U)	354
Figure 13.1 Forest plot of all studies that reported a CV for CRP, split by Risk of Bias	364
Figure 13.2 Forest plot of all studies that reported a CV for CRP, split by setting ..	365
Figure 13.3 Forest plot of all studies that reported a CV for CRP, split by long (L) or short (S) term measurement of variation	366
Figure 13.4 Forest plot of all studies that reported a CV for CRP, split by unit of measurement	367
Figure 13.5 Forest plot of all studies that reported a CV for hsCRP, split by Risk of Bias	368
Figure 13.6 Forest plot of all studies that reported a CV for hsCRP, split by setting	369
Figure 13.7 Forest plot of all studies that reported a CV for hsCRP, split by short (S) or long (L) term measurement of variation	370
Figure 13.8 Forest plot of all studies that reported an ICC for CRP, split by Risk of Bias, with meta-analysis.....	371
Figure 13.9 Forest plot of all studies that reported an ICC for hsCRP, split by Risk of Bias, with meta-analysis.....	372

Figure 14.1 Forest plot of within-individual CV of FEV1 by setting, all studies using units of Litres.	378
Figure 14.2 Forest plot of within-individual CV of FVC by setting, all studies using units of Litres.	379
Figure 14.3 Forest plot of within-individual CV of FEV1 by within or between-day, all studies using units of Litres.	380
Figure 14.4 Forest plot of within-individual CV of FVC by within or between-day, all studies using units of Litres.	381
Figure 14.5 Forest plot of within-individual CV of FEV1 by unit.....	382
Figure 14.6 Forest plot of within-individual CV of FVC by unit	383
Figure 14.7 Forest plot of within-individual CV of FEV1 by highest risk of bias score, all studies using units of Litres	384
Figure 14.8 Forest plot of within-individual CV of FVC by highest risk of bias score, all studies using units of Litres	385
Figure 14.9 Forest plot of within-individual SD of FEV1 by health status, all studies using units of Litres	386
Figure 14.10 Forest plot of within-individual SD of FVC by health status, all studies using units of Litres	387
Figure 16.1 CVT by region	404
Figure 16.2 CVT by sex.....	404
Figure 16.3 CVT by age	405
Figure 16.4 CVT by ethnicity	405
Figure 16.5 CVT by Townsend deprivation score	406
Figure 16.6 CVT by BMI.....	406
Figure 16.7 CVT by comorbidities	407

Figure 16.8 CVT by alcohol consumption.....	407
Figure 16.9 CVT by smoker status	408
Figure 16.10 CVT by year	411
Figure 16.11 CVT by quarter	411
Figure 16.12 CVT by Number of measurements	412
Figure 16.13 CVT by Mean	413
Figure 16.14 CVT by days between measurements.....	413
Figure 16.15 CVT by Days between measurements stratified by diabetic status...	414
Figure 16.16 CVT by unit of measurement.....	414
Figure 16.17 CVT by arithmetic calculation methods	415
Figure 16.18 CVT by mean	415
Figure 16.19 CVT by mean (log-transformed data)	416
Figure 17.1 CVT by region	419
Figure 17.2 CVT by sex.....	419
Figure 17.3 CVT by age	420
Figure 17.4 CVT by ethnicity	420
Figure 17.5 CVT by Townsend deprivation score	421
Figure 17.6 CVT by BMI.....	421
Figure 17.7 CVT by comorbidities	422
Figure 17.8 CVT by alcohol consumption.....	422
Figure 17.9 CVT by smoker status	423
Figure 17.10 CVT by year	428
Figure 17.11 CVT by quarter	429
Figure 17.12 CVT by Number of measurements	429
Figure 17.13 CVT by days between measurements.....	430

Figure 17.14 CVT by arithmetic calculation methods (using non-log adjusted data) and linear regression (using non-log adjusted data)	430
Figure 17.15 CVT by patient mean (non-log adjusted data).....	431
Figure 18.1 CVT by region for FEV1	434
Figure 18.2 CVT by region for FEV1	434
Figure 18.3 CVT by sex for FEV1	435
Figure 18.4 CVT by sex for FVC	435
Figure 18.5 CVT by age for FEV1	436
Figure 18.6 CVT by age for FVC	436
Figure 18.7 CVT by ethnicity for FEV1	437
Figure 18.8 CVT by ethnicity for FVC.....	437
Figure 18.9 CVT by Townsend deprivation score for FEV1	438
Figure 18.10 CVT by Townsend deprivation score for FVC	438
Figure 18.11 CVT by BMI for FEV1.....	439
Figure 18.12 CVT by BMI for FVC.....	439
Figure 18.13 CVT by comorbidities for FEV1	440
Figure 18.14 CVT by comorbidities for FVC.....	440
Figure 18.15 CVT by alcohol consumption for FEV1	441
Figure 18.16 CVT by alcohol consumption for FVC	441
Figure 18.17 CVT by smoker status for FEV1.....	442
Figure 18.18 CVT by smoker status for FVC.....	442
Figure 18.19 Heat map including time between measurements for long-term FEV1	449
Figure 18.20 Heat map including time between measurements for long-term FVC	449
Figure 18.21 CVT by year for FEV1	450

Figure 18.22 CVT by year for FVC	450
Figure 18.23 CVT by quarter for FEV1	451
Figure 18.24 CVT by quarter for FVC	451
Figure 18.25 CVT by Number of measurements for FEV1	452
Figure 18.26 CVT by Number of measurements for FVC	452
Figure 18.27 CVT by mean for FEV1	453
Figure 18.28 CVT by Mean for FVC	453
Figure 18.29 CVT by days between measurements for FEV1	454
Figure 18.30 CVT by days between measurements for FVC	454
Figure 18.31 CVT by unit of measurement for FEV1	455
Figure 18.32 CVT by unit of measurement for FVC	455
Figure 18.33 CVT by arithmetic calculation methods for FEV1	456
Figure 18.34 CVT by arithmetic calculation methods for FVC	456
Figure 18.35 CVT by mean short and long-term data for FEV1	457
Figure 18.36 CVT by mean short and long-term data for FVC	457
Figure 18.37 CVT by respiratory status for short and long-term data for FEV1 and FVC	458
Figure 19.1 CVT by region for PHQ-9	460
Figure 19.2 CVT by sex for PHQ-9	460
Figure 19.3 CVT by age for PHQ-9	461
Figure 19.4 CVT by ethnicity for PHQ-9	461
Figure 19.5 CVT by BMI for PHQ-9	462
Figure 19.6 CVT by comorbidities for PHQ-9	462
Figure 19.7 CVT by alcohol consumption for PHQ-9	463
Figure 19.8 CVT by smoker status for PHQ-9	463

Figure 19.9 Heat map including time between measurements for PHQ-9, interval between measures in days.....	465
Figure 19.10 CVT by year for PHQ-9	465
Figure 19.11 CVT by quarter for PHQ-9.....	466
Figure 19.12 CVT by Number of measurements for PHQ-9.....	466
Figure 19.13 CVT by Mean for PHQ-9	467
Figure 19.14 CVT by mean patient score for HADS.....	467
Figure 19.15 CVT by mean patient score for BDI.....	468
Figure 19.16 CVT by days between measurements for PHQ-9	468
Figure 19.17 CVT by arithmetic calculation methods for PHQ-9	469
Figure 19.18 CVT by log-transformed data for PHQ-9	469

TABLES

Table 1.1 Categories of clinical measurements.....	34
Table 1.2 Terminology and symbols used to describe biological variation.	47
Table 1.3 The intraindividual coefficients of variation (CVI) reported for selected laboratory tests.....	68
Table 1.4 Measures of variation and their equations or definitions.	69
Table 2.1 Variables extracted from studies	81
Table 2.2 Summary statistics of study characteristics of included studies	86
Table 2.3 Summary of results of variation measures of primary study population ...	87
Table 2.4 Summary statistics of reported CV reporting in units of %, split by health status.....	90
Table 3.1 Variables extracted from studies	108
Table 3.2 Summary statistics of study characteristics and results of included studies	111
Table 3.3 Highest risk of bias with questions B4 and B5 excluded.	117
Table 4.1 Characteristics of studies included (Paediatric populations).....	133
Table 4.2 Characteristics of studies included (Adult populations).	135
Table 5.1 Variables extracted from studies	159
Table 5.2 Depression measurement instruments analysed for variation in the included studies.	163
Table 5.3 Health status of patients included in studies.....	163
Table 6.1 Variables extracted from database.....	183
Table 6.2 Sociodemographic data and main univariable subgroup analyses.....	190
Table 7.1 Variables extracted from database.....	208
Table 7.2 Sociodemographic data and main univariable subgroup analyses.....	212

Table 8.1 NICE guidelines for categorising severity of airflow obstruction.	226
Table 8.2 Variables extracted from database.	230
Table 8.3 Sociodemographic data and main univariable subgroup analyses for FEV1 and FVC.	236
Table 9.1 Variables extracted from database.	259
Table 9.2 Sociodemographic data and CV of subgroups	265
Table 10.1 Median CVTs of systematic reviews.	288
Table 10.2 Summary of IMRD/CPRD studies of variation.	299
Table 12.1 Databases searched	333
Table 12.2 Risk of Bias form, Part A: Elements of Research Question	334
Table 12.3 Risk of Bias form, Part B: Risk of bias rating	336
Table 12.4 Included studies and characteristics.	340
Table 12.5 Characteristics of studies (continued)	342
Table 12.6 Risk of bias estimates for each included paper, with questions 4 and 5 excluded.	345
Table 13.1 Study characteristics detail.	358
Table 13.2 Study characteristics detail (continued) Characteristics of studies included.	360
Table 13.3 Risk of bias scores Maximum risk of bias score for each study with questions B4 and B5 excluded, and lowest BIVAC score (for studies that report a CV)	362
Table 14.1 Variables extracted from studies	377
Table 15.1 Databases searched	391
Table 16.1 Other subgroup analyses	402
Table 16.2 Sensitivity analyses	409

Table 16.3 CVT by quarter	412
Table 17.1 Other subgroup analyses	417
Table 17.2 Sensitivity analyses	425
Table 17.3 CVT by quarter	428
Table 18.1 Subgroup analyses for FEV1.....	432
Table 18.2 Subgroup analysis for FVC.....	433
Table 18.3 Sensitivity analyses for FEV1	443
Table 18.4 Sensitivity analyses for FVC.....	444
Table 18.5 Results for long-term data (i.e. all measurements included) for FEV1 .	447
Table 18.6 Results for long-term data (i.e. all measurements included) for FVC ...	448
Table 19.1 Other subgroup analyses.	459
Table 19.2 Sensitivity analyses.	464

Abbreviations

BDI	Beck's Depression Inventory
BMI	Body Mass Index
CKD	Chronic kidney disease
COSMIN	Consensus-based standards for the selection of health measurement Instruments
COPD	Chronic obstructive pulmonary disease
CPRD	The Clinical Practice Research Datalink
CRP	C-reactive protein
CV	Coefficient of variation
CV _A	Analytical coefficient of variation
CV _I	Within-individual biological coefficient of variation
CV _T	Total measured within-individual coefficient of variation
DM	Diabetes mellitus
EFLM	European Federation of Clinical Chemistry and Laboratory Medicine
EPR	Electronic Primary Care Record
FEV1	Forced expiratory volume in one second
FVC	Forced vital capacity
HADS	Hospital Anxiety and Depression Scale
HbA1c	Haemoglobin A1c
ICC	Intraclass correlation coefficient
IEC	International Expert Committee
IFCC	International Federation of Clinical Chemistry and Laboratory Medicine
IMRD	IQVIA Medical Research Database
MAOi	Monoamine oxidase inhibitor
MCID	Minimal Clinically Important Difference
MCT	Meaningful Change Threshold
NGSP	National Glycohemoglobin Standardization Program
NICE	National Institute for Health and Care Excellence
PHQ-2	Patient Health Questionnaire-2
PHQ-9	Patient Health Questionnaire-9
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PROM	Patient Reported Outcome Measure
QOF	Quality Outcomes Framework
SCID-5-CV	Structured Clinical Interview for DSM-V (Clinical version)
SD	Standard Deviation
SSRI	Serotonin reuptake inhibitor
STROBE	Strengthening the Reporting of Observational studies in Epidemiology
TCAD	Tricyclic anti-depressant
THIN	The Health Information Network
VIM	Variation Independent of the Mean
WHO	World Health Organization

1 Chapter one: Introduction

1.1 Clinical measurement and within-individual measurement variability

Clinical measurement is fundamental to clinical practice. Categories of clinical measurement include: laboratory tests such as glycosated haemoglobin HbA1c to evaluate glycaemic control and CRP as a marker of inflammation; clinical assessment such as measuring blood pressure, weight or respiratory function; structured questionnaires such as pain or depression symptom scores, or cognitive tests.

Clinical measurement is vital in the diagnosis, management and determination of prognosis of disease, and changes in treatment depend on the values generated by clinical measurement. These measurements vary due to chance and non-chance factors. However, the results of most clinical measurements are reported as a single value and these do not reflect the uncertainty and variability underlying these measurements [1]. Biological processes vary within an individual over time and much of this is a result of biological variability due to chance [2]. Furthermore, many clinical measurements are continuous variables, whereas diagnosis is usually binary or categorical (i.e. yes/no, or categorising patients based on test results). For example, blood pressure, measured in mmHg, is a continuous variable, while the diagnosis of high blood pressure, hypertension, is binary, i.e. hypertensive or not, and the National Institute of Clinical Excellence (NICE) guidelines recommend beginning treatment with antihypertensives when blood pressure is measured as “160/100 mmHg or higher but less than 180/120 mmHg and subsequent ABPM [ambulatory

blood pressure measurement] daytime average or HBPM [Home blood pressure measurement] average blood pressure of 150/95 mmHg or higher”[3].

This thesis aims to 1) summarise published data (using systematic reviews) of within-individual variation for selected laboratory and non-laboratory clinical measurements and 2) estimate and describe within-individual variation in selected laboratory and non-laboratory clinical measurements using anonymised data from electronic primary care records (EPRs) from the IQVIA Medical Research Database (IMRD) and Clinical Practice Research Datalink (CPRD).

1.1.1 What is measured in clinical settings?

Most clinical measurements fall into one of six categories which for the purposes of this thesis are defined as follows: laboratory tests; imaging tests; clinical characteristics; functional/physiological measurements; symptom scores; clinical signs (see figure 1.1 and table 1.1). Note there may be some overlap between these categories. For example, estimated glomerular filtration rate (eGFR) which is used to measure kidney function could be considered both a functional test and a laboratory test, or cognitive testing may be considered a functional test or a symptom score. Nevertheless, this categorisation is useful to conceptualise the variety of testing that takes place in clinical practice.

Table 1.1 Categories of clinical measurements

- a. Laboratory testing includes haematological and biochemical testing, as well as histology, cytology and infectious disease testing. Examples include neutrophil count, blood glucose, estimated glomerular filtration rate (eGFR), antibody titre tests and bacteriology sensitivity tests. The results of laboratory tests are mostly presented numerically, but some are subjective and non-numerical, such as histological diagnosis and some are binary e.g. presence/absence of disease.
- b. Imaging tests such as x-ray, computed tomography (CT), ultrasound and magnetic resonance imaging (MRI) are often interpreted subjectively and non-numerically [4]. However, some results are reported numerically, such as measures of lesion size, measures of heart function with echocardiography and measures of brain function with functional MRI [5], as well as scoring systems such as the modified Alvarado score for sonographical features of acute appendicitis [6] or the Brasfield or Wisconsin radiological scoring systems for cystic fibrosis [7]. Although variation in imaging is an important subject, with variation within individuals, between individuals and between observers, imaging is rarely carried out in primary care practice and so will not be considered in this thesis.
- c. Clinical characteristics are measurements that describe a physical characteristic of a patient. Examples include height, weight, body mass index, waist circumference, urine output, blood pressure and heart rate. These are often reported as continuous numerical variables.
- d. Functional measurements describe the extent of the ability of a patient or body system to function optimally. Examples include spirometry, hand grip strength and visual acuity. These are also often reported as continuous numerical variables.
- e. Symptom scores attempt to describe subjective experiences in a numerical fashion. Examples include pain scores [8], the Sodergren symptom-based severity score for haemorrhoidal disease [9], cognitive function scores such as the mini-mental state examination [10] and depression scores such as the Patient-Health Questionnaire-9 (PHQ-9) [11]. These are often reported as numerical scales.
- f. Clinical signs include mucous membrane colour, skin lesions, liver enlargement and chest sounds on auscultation. These are usually reported non-numerically and subjectively, or are at best binary or semi-quantified (e.g. clinical scores), so variation within an individual is hard to assess. For this reason, clinical signs will not be considered in this thesis.

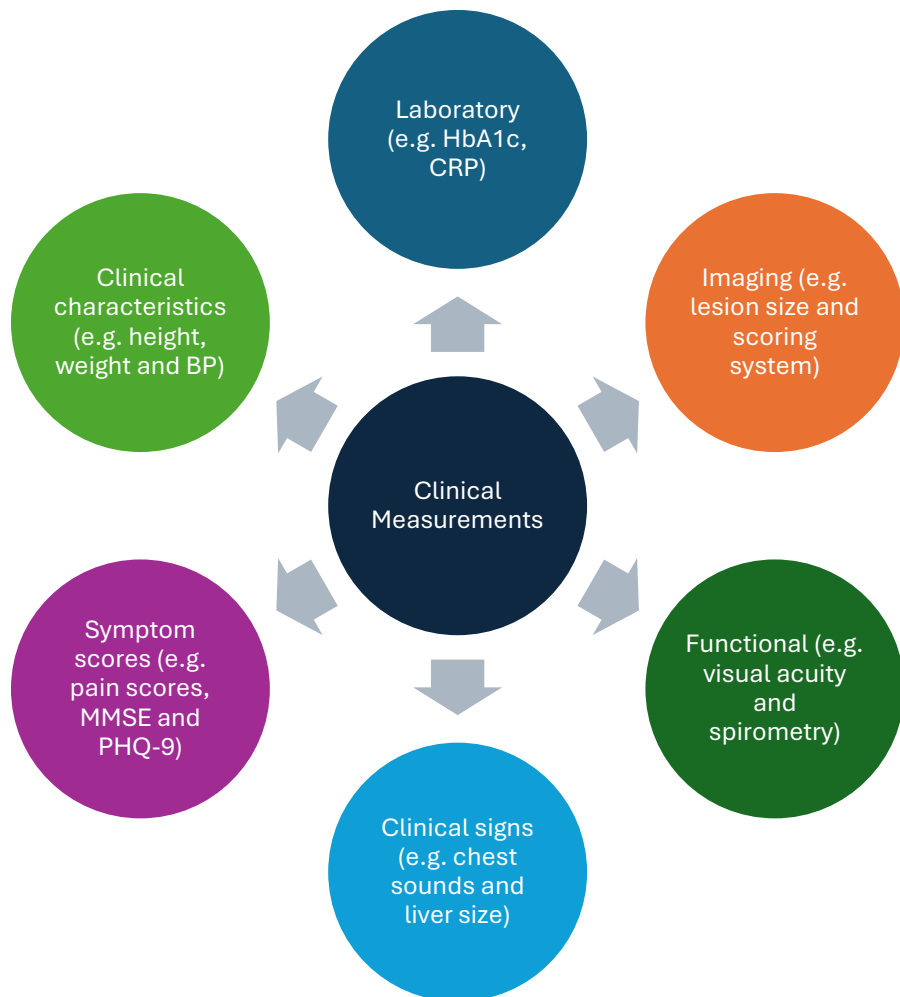


Figure 1.1 Categories of clinical measurements with examples

The terms used in measurement need carefully defining. For the purposes of this thesis, the biological process or status being measured will be referred to as the measurand. Examples include blood glucose and systolic blood pressure. A measurement or result is the number or value generated by the test or clinical procedure. Examples include a glucose level in mmol/l or a blood pressure measurement in mmHg. The act of performing the measurement will be referred to as the test or clinical procedure, or by the specific name of the test, such as blood glucose test or blood pressure test.

1.1.2 What are clinical measurements used for?

Tests are used for diagnosis and monitoring of disease, screening of apparently healthy but at-risk patients, and estimating prognosis. There are many reasons why a practitioner may order a test in a patient. These have been categorised into diagnostic factors, therapeutic and prognostic factors, patient-related factors, doctor-related factors and policy or organisational factors [12]. Diagnostic factors include modifying a pre-test probability, making a definitive diagnosis, or deciding when to refer for secondary care. Therapeutic and prognostic factors include making treatment decisions, monitoring treatment effects and predicting the clinical course of a condition. Patient factors can influence the decision to order a test, such as safety of the test, impact of diagnosis or lack of diagnosis and desire for reassurance. Physician factors also influence the decision to order a test such as clinical experience and self-confidence, ability to tolerate uncertainty and over-estimation of the value of tests [12, 13]. Policy or organisational factors include practice size, ease of ordering tests, method of physician remuneration and clinical guidelines.

If physicians are using testing to reduce their diagnostic or prognostic uncertainty, it is important they are aware of the level of uncertainty inherent in the tests they order. Uncertainty can arise because of measurement variation and biological variation.

1.2 What are the sources of variability in a measurement?

Figure 1.2 illustrates the factors that influence a clinical measurement. Firstly, biological and pathological changes influence the “true” value of the measurand (which cannot be observed). The measurement process then introduces further variability, leading to a measured biological status (a “test result”). Finally, this test is interpreted by a clinician, leading to a clinical decision or judgement regarding diagnosis, treatment and/or prognosis. The various terms used in the literature to describe biological variation are given in table 1.2 and fig 1.2. These will be discussed in more detail below. In brief, pathological variation refers to variability due to the derangement of body functions caused by disease. Biological variation refers to the longitudinal fluctuations in biological values which can be systematic and potentially predictable, or due to chance. Pre-analytical variation involves the variability introduced by differences in patient preparation, sampling procedure and sample handling. Analytical variation is introduced at the level of the analysis, beginning from when a sample starts to be prepared for analysis, and ending when a result is generated. Post-analytical variation refers to the interpretation and use of these results.

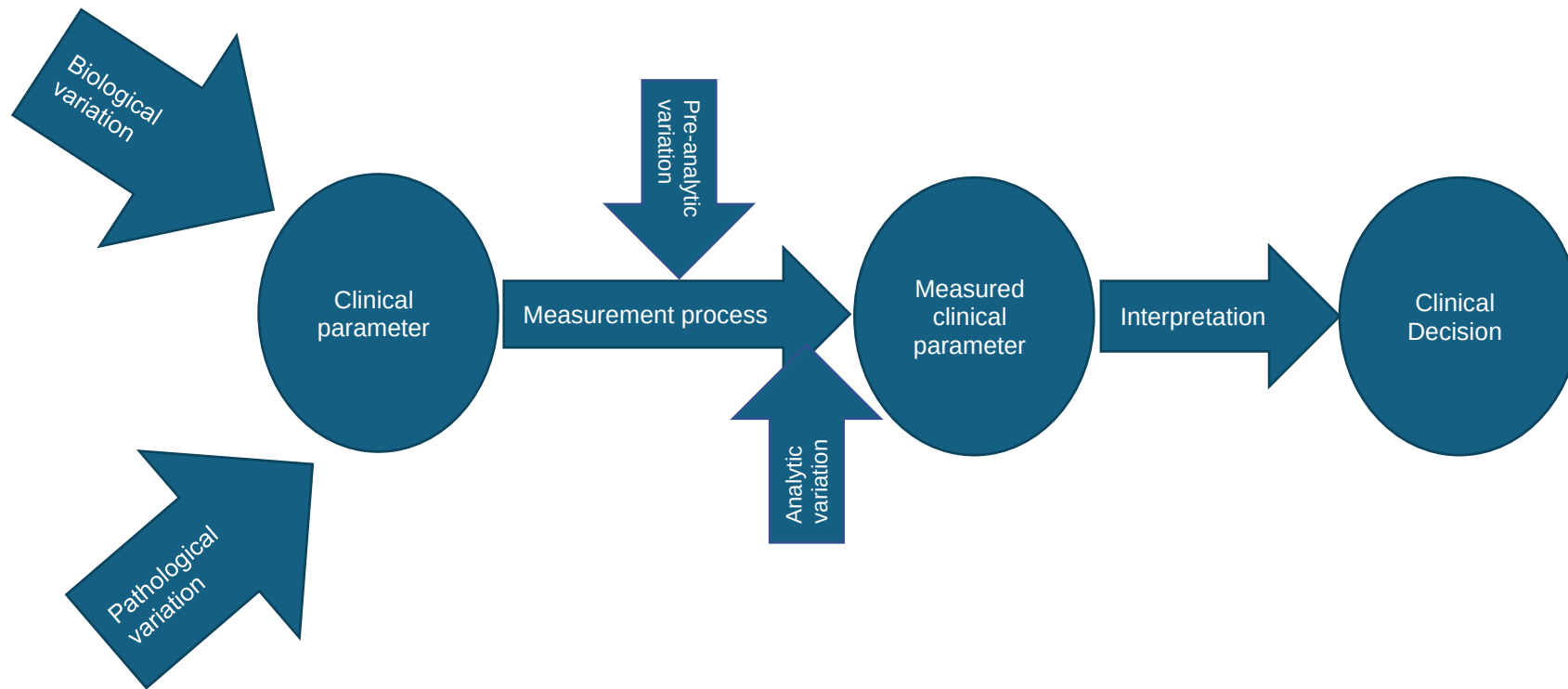


Figure 1.2 Sources of within-individual variability in clinical measurements and decision-making

1.2.1 Biological variation

Biological variation is the term used to describe variability measured over time, and refers to the component of total variability (that is the variability in a measurement from all sources) due to biological processes. Biological variation can be systematic and to an extent predictable, or can be due to chance. Even if analytical precision (the closeness of results performed on identical samples) was perfect, there will be differences in the results of a measurement taken at different time points in the same individual[14]. Homeostasis is the concept that the body aims to keep a constant internal environment in order to function optimally. Most biological processes and statuses measured in laboratories have homeostatic set points in individuals, that is the physiological value around which the range of the process or status will vary. As external or internal factors move the process or status away from the set point, homeostatic mechanisms kick in, for example release of insulin if glucose increases, or sweating if body temperature increases, to move the value back towards the set point. Variation between individuals (between-individual variation) however represents different homeostatic set points between individuals[1, 15].

1.2.2 Pathological variation

Pathological variation is the variation in a disease process or the measured results due to the disease itself. Pathological variation can be thought of in two different ways, the variation due to a material change in the disease or an increased level of variability due to the disease. Examples of a material change in disease which might

appear as variation in measured results include disease progression (e.g. decline of pulmonary function in COPD), natural improvement in the disease process (e.g. resolution of an infection due to the body's immune system) or a result of interventions (e.g. antibiotics to treat bacterial infections). Pathological variation often occurs in a single direction over time for a given disease, for example a decline in kidney function (decreasing eGFR), or a decrease in lung function with chronic obstructive pulmonary disease (COPD) documented by spirometry measures. However, some pathological variation is temporary, for example in successfully treated or self-limiting diseases, or will move to vary around a new mean in a well-controlled chronic disease, such as HbA1c in diabetes mellitus.

Some pathologies increase the variability of a biological process or status. For example, blood glucose variability is exaggerated in type I diabetes. Blood glucose levels are tightly controlled by insulin secretion. An increase in blood glucose causes an increase in insulin secretion (and a reduction in hyperglycaemic hormones such as glucagon), which reduces the blood glucose levels. Conversely, a reduction in blood glucose, for example due to exercise, reduces the insulin secretion (and increases hyperglycaemic hormones) which increases blood glucose levels. Since in diabetes mellitus insulin secretion is reduced, these minute-to-minute changes in insulin secretion to counteract changes in blood glucose are blunted [16], leading to larger swings in blood glucose, i.e. higher variation. Increased variation of spirometry measures Forced Expiratory Volume 1 second (FEV1) and Forced Vital Capacity (FVC) are seen in asthma [17].

It is important to attempt to distinguish between these two aspects of pathological variation, a material change in the underlying disease and increased variability in the

disease caused by pathology. The former is the “signal” on which useful decisions can be made, and the latter is the “noise” which complicates decision-making.

1.2.3 Pre-analytical variability

Pre-analytical variation is introduced during the pre-analytical phase of laboratory testing, which begins even before a sample is collected, and ends when the sample begins to be prepared. It is thought to be a major cause of laboratory error, partly because it relies on personnel outside the direct supervision of the laboratory setting, such as nurses and general practitioners [18]. However, there are numerous other factors besides human error in the pre-analytic phase that can cause variability.

Causes of pre-analytic variation can be divided into patient factors, collection factors and post-collection factors. Patient factors include exercise, fasting/starvation, altitude, stimulants and drugs [15, 19]. Collection factors include posture (the measured value of some analytes can vary up to 10% if taken from a recumbent compared to a standing posture), prolonged use of tourniquet, type of sample (e.g. capillary versus venous blood sample, serum versus plasma), the type of anticoagulant used, the ratio of the volume of sample to anticoagulant (e.g. insufficient sample) and haemolysis due to problematic venepuncture and contamination with exogenous fluids [15, 19, 20]. Post-collection factors include transport time, centrifugation time and speed, and storage (e.g. temperature, exposure to light, contact with air) [15, 18, 20]. Different tests can be more prone to some of these effects than others. For example, cholesterol may increase by up to 18% when changing from a supine to a standing posture [19] and packed cell volume (PCV), haemoglobin and neutrophil count all varied significantly in samples

collected when standing or supine, while monocytes, mean corpuscle volume (MCV) and mean corpuscle haemoglobin (MCH) did not [21].

Non-laboratory tests can also be affected by pre-analytical variation. For example, blood pressure can be affected by the setting in which it is measured. Anxiety can lead to the release of catecholamines and activate the sympathetic nervous system so blood pressure measured in clinic may be higher than at home – so called white coat hypertension [22]. Pain scores can be influenced by a host of psychological and physiological factors. For example, the pain score associated with intravitreal injection was found to be lower in patients who had received a previous intravitreal injection that led to an improvement in vision [23].

Note there is a lack of agreement on the definition of pre-analytical variation, and what should be included in this term [24] [25]. In this thesis I will consider biological variation, whether it be random or systematic, to be separate from pre-analytical variation. However, as discussed below, the variation data obtained from the IMRD and CPRD datasets will be total measurement variation including pre-analytical, analytical and biological variation components.

1.2.4 Analytical variation

The analytical phase of testing in laboratory terms occurs when a patient sample begins to be prepared for testing. Analytical variation can be random or systematic. The degree of systematic variation affects the bias in a test, while random variation affects precision (see fig 1.3).

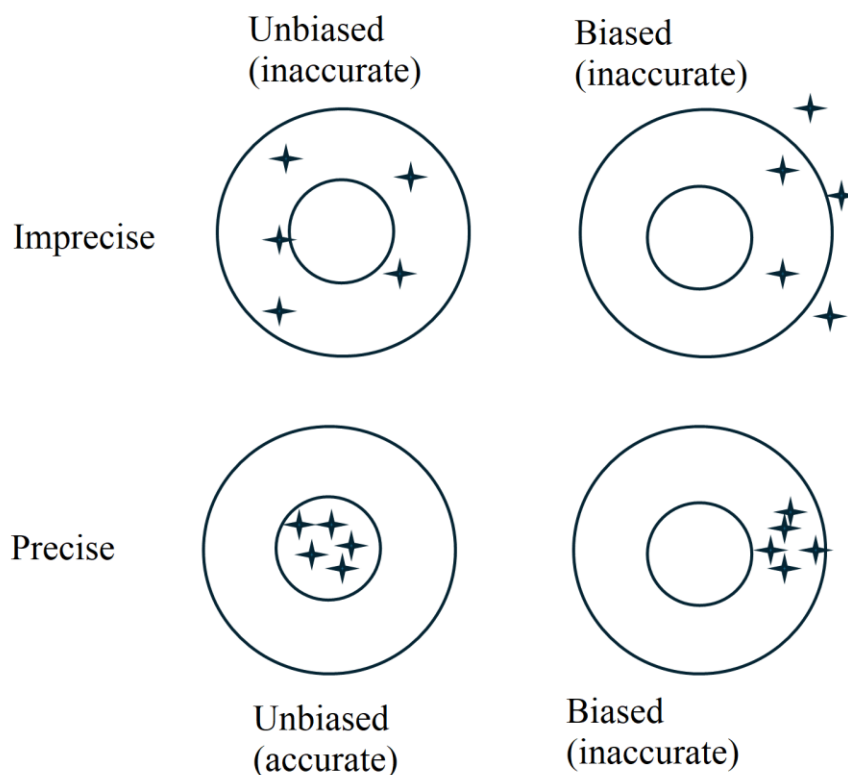


Figure 1.3 Illustration of bias, precision and accuracy, using average distance to target centre of five shots

Precision has been defined as the, “closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions” [26]. Bias has been defined as the, “systematic difference between the results obtained by the lab’s test method and the results obtained from an accepted reference method”[27]. Random variation arises in part due to the measurement device or analytical system and the methodology, related for example to variations in the volume of sample or reagent in the pipette, as well as problems with immunochemical methods, for example heterogeneity in the analyte causing underestimation due to over-specific monoclonal antibodies, cross reactivity and autoantibodies against the analytes [28].

Bias may be caused by variations between or within laboratories regarding equipment used or batches of reagents [29, 30], by individual technicians not following standard operating procedures [15, 31] and by equipment malfunction [31].

Bias and imprecision can be combined to make a more intuitively comprehensible measurement known as the Total Analytical Error (TAE) [26]. An alternative model [32] uses the concept of Measurement Uncertainty (MU) in which measurements are considered in relation to the belief placed on their accuracy, which is expressed as an uncertainty. The latter model does not rely on the concept of a true value, unlike the former model, and as such these two paradigms seem incompatible. For the purposes of monitoring, the acceptable analytical variation is usually set at <0.5 of the intra-individual biological variation [32]. Analytical variation is often expressed as the coefficient of variation (CV_A).

Non-laboratory measures can also be affected by analytical error, usually due to imprecision in the measurement device. Intrinsic imprecision may be exacerbated by poor calibration and maintenance, for example in-clinic spirometry machines [33], but will always be present to some extent even in the most accurate and well-maintained devices. In veterinary medicine, a hoof tester is used to apply pressure to various parts of a horse's hoof to test for pain. The amount of force applied has been shown to vary markedly within and between individuals [34]. Intra- and inter-observer reproducibility is a well-understood problem in imaging. Standardisation protocols can reduce this error but do not eliminate it [35, 36].

Some clinical conditions are assessed using questionnaires. This is uncommon in most clinical areas, although some think it is underused [37]. However, some conditions are hard to assess with laboratory or objective clinical measures, and so

questionnaires are more commonly used, such as in assessment of depression and dementia. The use of test-retest reliability estimates can be considered a way of estimating the analytical variability of a questionnaire, and this is performed commonly, (in contrast to the within-individual variation of questionnaires which seems to have been little considered).

1.2.5 Post-analytical variation (Interpretation and clinical decision-making)

Post-analytical variation describes the variation in decisions made after interpreting a clinical measurement. Variation in clinicians receiving reports and acting on them in a timely manner come into this category. Also the decisions made based on the laboratory result generated are variable, and errors related to variation can also appear in decision limits such as reference ranges and clinical guidelines [38, 39].

1.2.6 Total within-individual variation

Within-individual biological variation (CV_I) is the combination of the systematic within-individual longitudinal changes such as circadian or seasonal changes of a biological measurand and the chance fluctuations around a homeostatic set-point, as described above, while analytical variation is the variation introduced by the act of measurement e.g. the variation inherent in using a piece of laboratory or clinical equipment (CV_A). The results within some of the published data on within-individual variation, and the results of tests in the IMRD and CPRD databases do not separate these components, and instead provide the within-individual total measurement variation (CV_T). These results represent the real-world data on which clinicians make

clinical decisions, and therefore they are the most relevant measure when considering the impact of within-individual variation on clinical practice[99].

1.3 Terminology

There are a number of different terms used in the literature to describe variation (see tables 1.2 and figure 1.4). Analytical variation is the term used to describe the component of variation due to the measurement process itself. Within-individual variation is the component of variation due to changes within a single individual. Between-individual variation refers to the variation of measurements between groups. This is a cross-sectional rather than a longitudinal variation, and is not considered in this thesis.

Although the term “within-subject” has been proposed to harmonise the terminology [40], the terms “within-individual variation” and “between-individual variation” are used throughout this thesis, as this is more consistent with the published literature. Within-individual biological variation is defined as “fluctuation of the concentration of constituents around their homeostatic set point,” or “variation within a single individual estimated as a pooled variation from a group of individuals,” or “variation within a single individual over time.” Between-individual biological variation is defined as “the differences between individual set points,” or “the variation between the central tendencies of a group of individuals.” [15, 40, 41]

The terms “variability” and “variation” appear to be used synonymously in the literature, but they have slightly different meanings. Variation is the set of observed differences while variability is the tendency of a system to generate differences [42].

In this thesis therefore I will use variation to refer to the observed patterns and results, while variability will be used to discuss underlying mechanisms.

Variation is often expressed as the Coefficient of Variation (CV) which is the standard deviation divided by the mean (See 1.2.3 below).

Table 1.2 Terminology and symbols used to describe biological variation.

Synonymous terminology for within-individual variation	Synonymous terminology for between-individual variation
Within-individual variation*	Between-individual variation*
Within-subject biological variation	Between-subject biological variation
Intraindividual biological variation	Interindividual biological variation
Within-subject variation	Between-subject variation
Intraindividual variation	Interindividual variation
Within subject CV	Between subject CV
Intraindividual CV	Interindividual CV
CV_i^*	CV_G^*
CV_T^*	CV_g
CV_{A+I}	CV_b
CV_i	
CV_w	
CV_W	
$CV_{\text{within-subject}}$	
$CV_{\text{biological}}$	
CV_b	

* = preferred terminology. After Simundic [40]. Note that CV_b has been used to refer to both “within-person variation” and “interindividual CV”

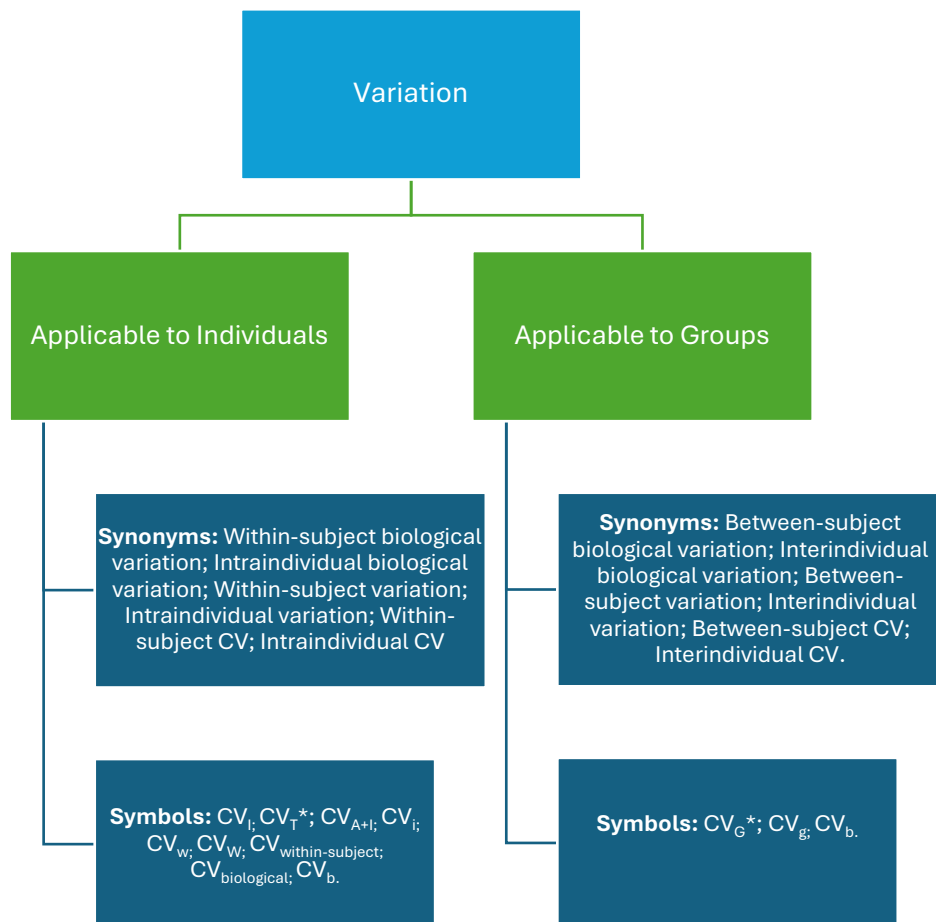


Figure 1.4 Terminology used to describe variation.

** = terminology used in this thesis.*

1.4 Why is within-individual variation a problem?

The variation in an individual's measurements over time, particularly where this variation is by chance and therefore not predictable, has important implications for diagnosis and treatment decisions during monitoring. For example, measured blood pressure may be above or below the true mean blood pressure. Some people will have blood pressure measured as high purely by chance, which will lead to them being misclassified as having hypertension. It has been shown that routine measurement of blood pressure in persons under 35 is more likely to misdiagnose than to diagnose hypertension correctly [43].

Clinical measurements are used in practice primarily for initial diagnosis, monitoring with a view to altering treatment, and prognosis. However, as mentioned previously, measurements are usually continuous, while diagnosis is usually binary, and treatment decisions are binary or categorical.

Four examples of tests which already have some published work on their level of variation - blood pressure, C-reactive protein, estimated Glomerular Filtration Rate and albumin:creatinine ratio - help illustrate the problems that can arise when within-individual variation is not considered.

1.4.1 Example 1: Blood pressure

Not taking into account within-individual measurement variation can lead to the clinically important errors of false positives and false-negatives [44]. NICE guidelines recommend that patients with stage 2 hypertension, defined as "Clinic blood pressure of 160/100 mmHg or higher but less than 180/120 mmHg and subsequent ABPM daytime average or HBPM average blood pressure of 150/95 mmHg or higher," (where ABPM stands for ambulatory blood pressure measurement and

HBPM stands for home blood pressure measurement) should begin anti-hypertensive treatment [3]. Although the recommendation for more than one measure of blood pressure to be taken reduces the influence of within-individual measurement variation, it does not eliminate it. There will be a proportion of individuals whose measured mean blood pressure will be above the guideline cut offs while their true mean blood pressure will be below the cut off. These patients will be falsely diagnosed with hypertension (false positive results), and will be treated unnecessarily and possibly harmfully with anti-hypertensive medication. Conversely, there will be a proportion of individuals whose measured mean blood pressure will fall below the guideline cut offs, while their true mean blood pressure will be above the cut offs. The patients with false negative results will not be treated with anti-hypertensive medication which may be harmful.

1.4.2 Example 2: C-reactive protein

Many guidelines give single figure cut off points for making decisions. As an example of a diagnostic decision with a single figure cut off, NICE guidelines on the use of antibiotics in patients with suspected pneumonia state, “Do not routinely offer antibiotic therapy if the C-reactive protein concentration is less than 20 mg/litre” [45]. C-reactive protein has a marked variation within individuals, and so a single measurement may not be sufficient to provide accurate diagnosis or clinical guidance to the physician [46, 47].

1.4.3 Example 3: Estimated glomerular filtration rate and albumin creatinine ratio

For monitoring and prognostic decisions in chronic kidney insufficiency, glomerular filtration rate and albumin creatinine ratio measurements are recommended in guidelines and reviews [48, 49]. But both have high within-individual measurement variation, meaning the “true” mean may put the patient in a different stage of the disease than the measured value [50, 51]. For example, the NICE guidelines state that patients who “have a GFR of less than 30mL/min/1.73m²” should be offered a renal ultrasound. Yet because of the high within-individual measurement variation of GFR, a sizeable proportion of patients with a true GFR greater than this cut off will have a measured GFR below the cut off (false positives) and will get an unnecessary ultrasound scan, causing mental distress and using up healthcare resources. Conversely a proportion of patients with a true GFR below this cut off will have a measured GFR above the cut off (false negatives) and will therefore not receive an ultrasound scan that may have been beneficial to their health.

1.5 Testing in primary care

Primary care is of vital importance in the delivery of healthcare. [52]. In this setting, the majority of the population’s preventative and curative health needs can be met [53]. Testing is a vital part of healthcare for diagnosis and monitoring, and therefore this thesis will examine the within-individual variation of testing within primary care.

1.5.1 How often are tests repeated in general practice?

Testing in general practice has increased markedly in recent years, with age- and sex-adjusted use increasing annually by 8.5%. One study found that in 2015/2016 patients had on average five tests per year, compared with 1.5 in 2000/2001 [54]. The Quality Outcome Framework (QOF) is a voluntary incentives and rewards scheme for GP practices in the UK, which among other indicators recommends certain tests to be performed for specific patients at specific times, for example TSH and creatinine in patients with liver disease every 9 months or HbA1c measurements in diabetics at least every twelve months [55]. Further guidance on testing and screening comes from The National Institute for Health and Care Excellence (NICE) such as recommendations for diagnosis and monitoring of renal disease or diabetes [56].

A pilot search of a sample of electronic patient records (EPRs) of GP practices in the IMRD database (ten GP practices, comprising 239,003 patients over a twenty year period) was performed to search for the frequency of performance of selected tests. The result is shown in figure 1.5. The data was then analysed to assess how often these selected tests were repeated in the same patient. This data is important for this thesis, since multiple tests in the same patient are needed to calculate within-individual variation. This data is shown in figure 1.6, presented as the median number of tests. However, the data was highly skewed, and some tests were performed dozens of times in the same individual.

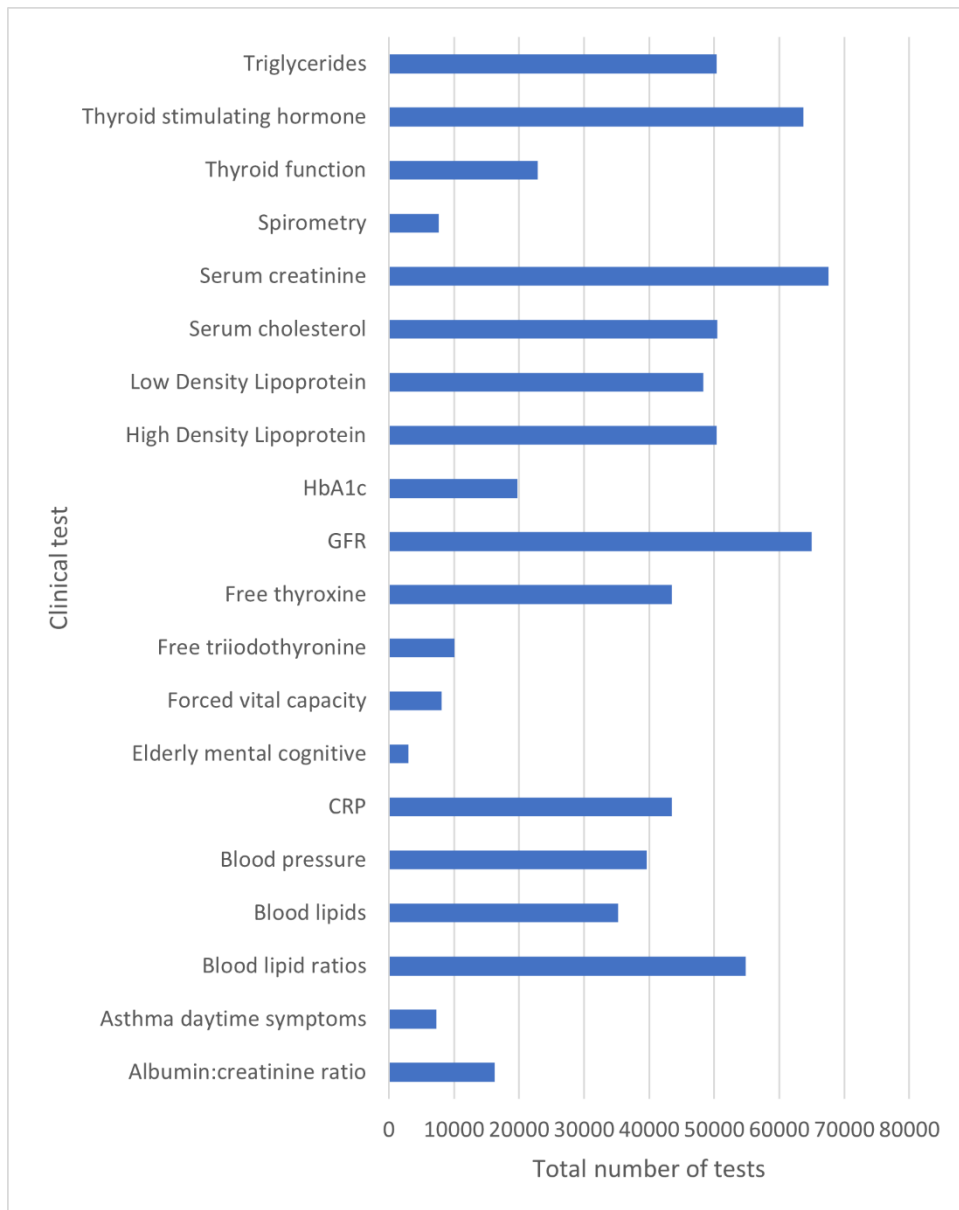


Figure 1.5 Frequency of selected tests performed in a sample of GP practices taken from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.

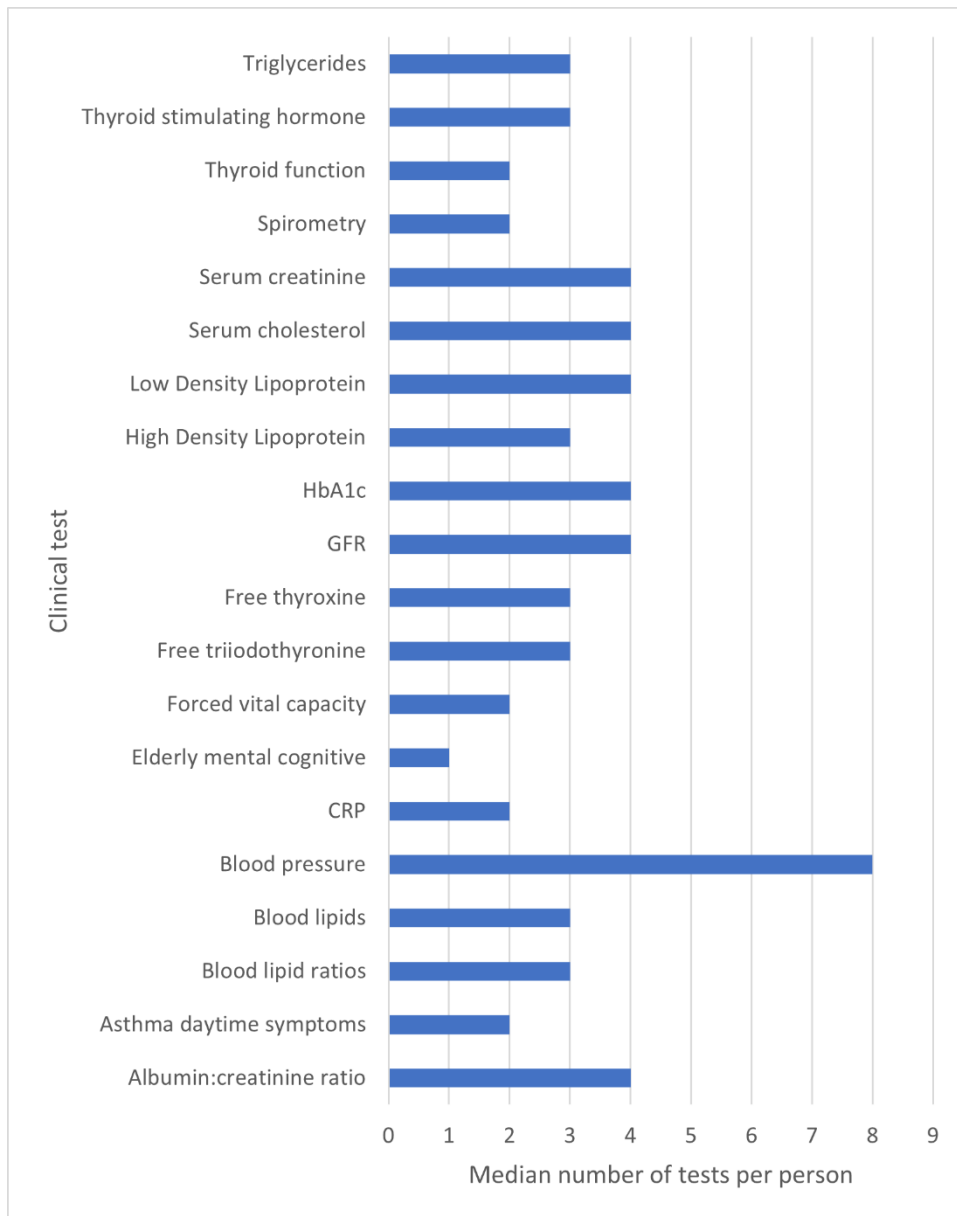


Figure 1.6 Median number of repeats in the same individual of a selection of tests performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.

Examples of laboratory and functional tests performed in primary care that are associated with a clinical decision are given in figs 1.7, 1.8 and 1.9.

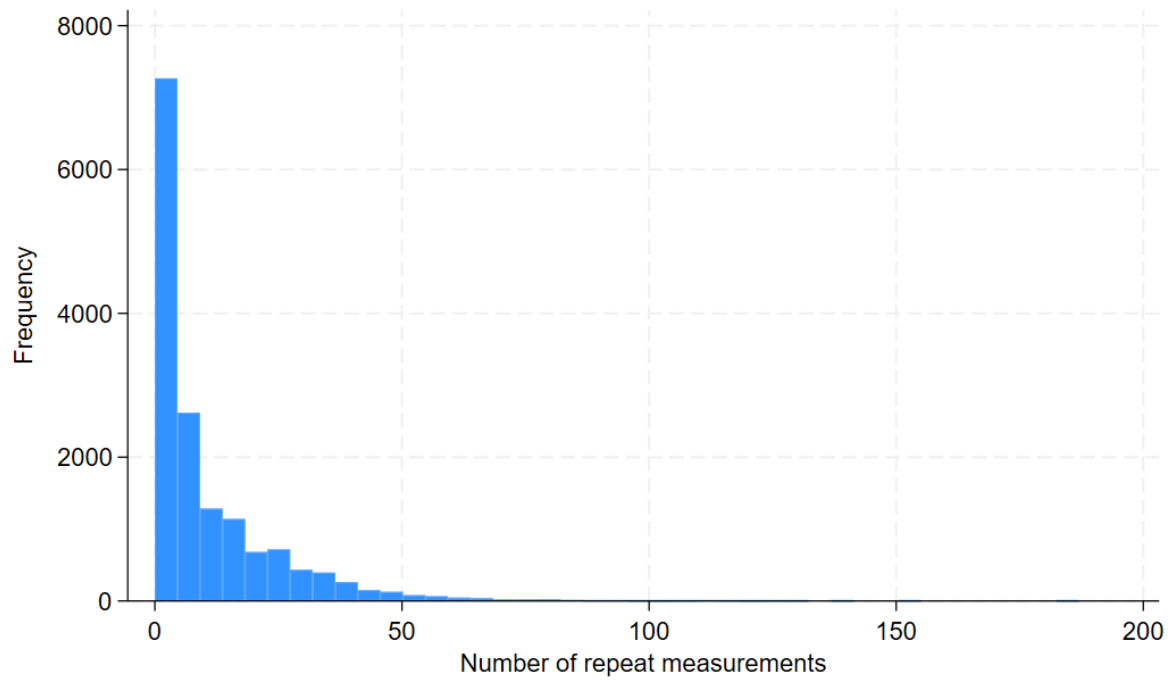


Figure 1.7 Distribution of number of repeated tests of HbA1c in the same individual performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.

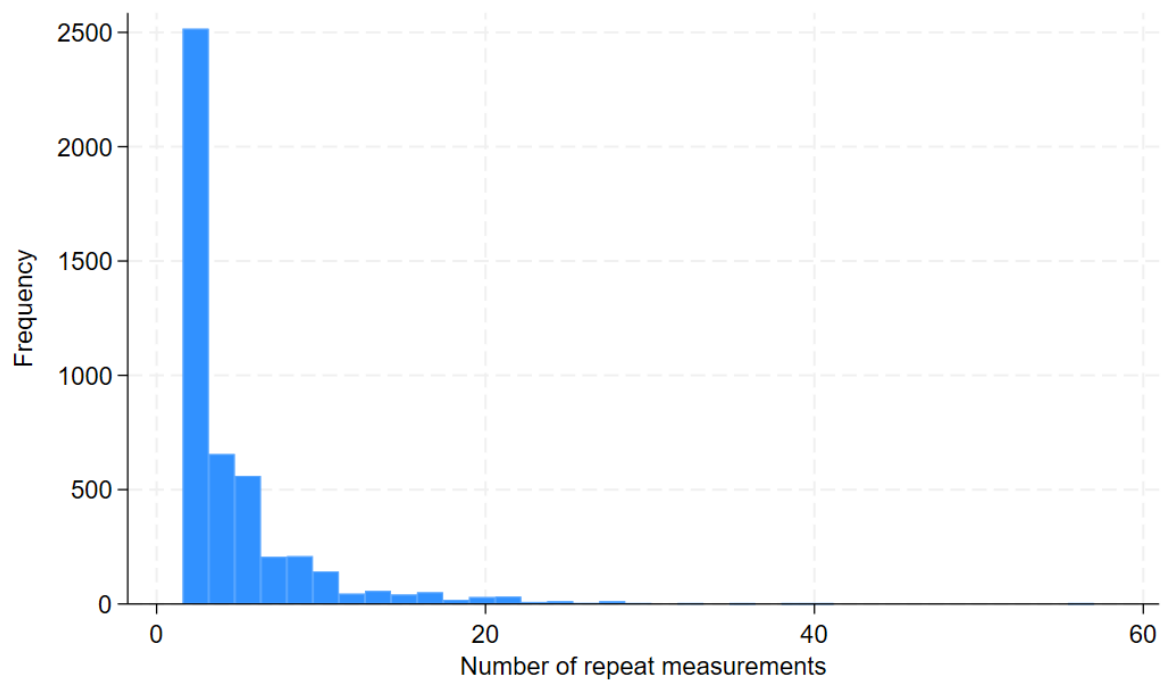


Figure 1.8 Distribution of number of repeated tests of forced vital capacity (FVC) in the same individual performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.

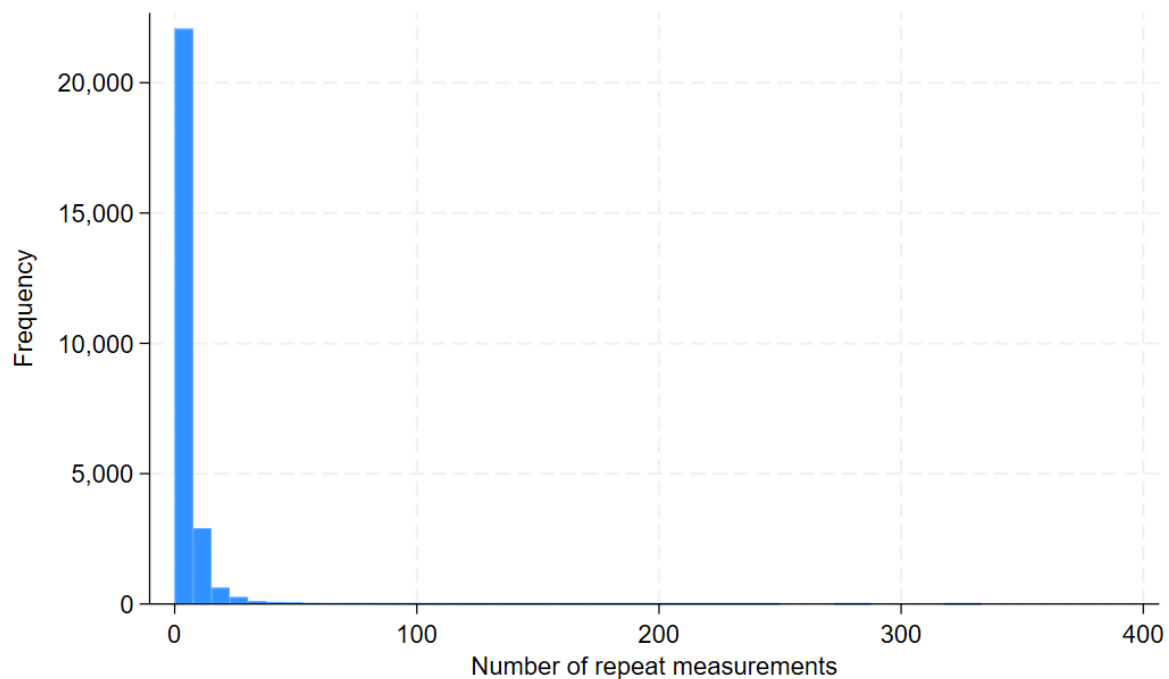


Figure 1.9 Distribution of number of repeated tests of C-reactive protein (CRP) in the same individual performed by a group of practices from the IMRD database. Data from a random sample of ten GP practices, comprising 239,003 patients over a twenty year period.

It can be seen that a number of tests are repeated frequently in the same individual, which is necessary (but not on its own sufficient) for them to be included in this study.

1.6 The effect of within-individual measurement variation on test results in practice

There are numerous guidelines and recommendations for diagnosing and monitoring diseases in clinical medicine. These usually rely primarily on the results of clinical measurements performed in practice or at a laboratory. However, as has

been shown, these results can vary markedly. Variation can be systematic and predictable, such as seasonal or circadian, but the largest part of the variation is often due to chance. Since acceptable analytical error of laboratory tests is usually set at <50% of the biological variation [32], the majority of the unpredictable variation over time is often chance biological variation [15].

1.6.1 The problems with reference ranges

When laboratory tests and most other clinical tests are reported and interpreted, a reference range is used to determine whether the results are likely to represent a normal situation, i.e. a healthy result, or abnormal. A good reference range “correctly includes most of the individuals with characteristics similar to the reference group and excludes the others” [57]. The majority of reference ranges in use are based on the central 95% of results from a healthy population [58]. These reference ranges are usually generated based on single samples from each individual in a population of interest. Many reference ranges are established in a population of healthy adults. Some reference ranges are calculated to take into account populations with specific characteristics, such as age, race, level of athletic ability etc. For example a reference range has been calculated for neonatal haematological values using the 5th to 95th percentile values of blood tests taken from neonates with minimal pathology [59], and a reference range has been produced for complete blood count in people over 50 years of age from single blood samples taken from volunteers [60]. A systematic review of reference values for right heart size and function in athletic individuals produced new reference ranges, since reference ranges for the general population could cause a mistaken assessment of the right heart and function in athletes [61]. C-reactive protein (CRP) has been

shown to have different reference intervals in people of black ethnicity compared to those of white and Hispanic ethnicities [62].

The problem with these reference ranges is that while they take into account between-individual variation in measurement values, they do not account for within-individual variation over time. This may or may not be important, depending on the degree of within-individual variation compared to the size of the reference range.

Note that reference ranges and test thresholds for decision making are similar concepts but not synonymous. For example, the reference range for TSH in adults based on the central 95% has been reported as 0.58-4.07 mU/l (although this can vary depending on genetic and environmental factors [63]. However, as noted below, the decision threshold for primary hypothyroidism is a TSH of 10mU/l [64].

It is also important to note that reference ranges and decision-making thresholds impose a cut-off point to a continuous measure, in contrast to a binary diagnostic test (for example presence or absence of a fracture on a radiograph).

1.6.2 Repeat testing

Repeated testing may be used to reduce the problem of within-individual variation, but its success depends both on variation and thresholds for decision-making. The number of tests required to get an accurate result that compensates for within-individual variation can be calculated. Cholesterol and CRP can be used for examples.

Fraser[65] uses the formula

$$n=(Z*[CV_a^2 + CV_b^2]^{1/2}/D)^2$$

to estimate the number of tests required to get an accurate result. N is the number of tests, CV_a is the analytical variation, CV_b is the within-individual biological variation, D is the closeness to the true homeostatic set point and Z is the Z-score appropriate for the probability (e.g. 1.96 for 95%). He assumed a CV_b (aka CV_i) of 6% and a CV_a of 3% for cholesterol, which meant that 2 samples were needed to get within 10% of the true value with a 95% probability. When the same method was applied to CRP, with a CV_a of 10% and a CV_b of 30%, the number of tests required was 38. And yet past guidelines from the Centers for Disease Control and Prevention and the American Heart Association stated that despite acknowledging the high within-individual biological variation of CRP, only two samples were needed to compensate for this variation to perform an accurate result [66].

1.7 The importance of within-individual measurement variation on decision-making in general practice

It has been estimated that 80% of a doctor's medical decisions are based on the results of laboratory testing [58]. Producing believable results and correctly interpreting them is vital to patient safety. Yet despite the knowledge and understanding, at least in some areas of medicine, that variation and uncertainty are part of all clinical measurements, this is still often poorly applied in practice. Tests results are usually reported as a single number, with no estimate of uncertainty, and often with inappropriate numbers of decimal points, misleading the clinician as to the accuracy of the result [1].

However, the applications of a test in clinical practice differs between tests, and different thresholds for decision-making between tests can mean that the variation of the measurand is more or less important, depending for example how far the threshold at which a decision is recommended is from the mean of the measurand. It is important to note, however, that clinical decisions are usually made taking into account a variety of factors such as the magnitude of the test result above the recommended threshold, clinician factors and patient factors. Nevertheless, the decisions of clinicians vary with a variety of factors including gender and belief in evidence-based medicine, making some clinicians more likely to adhere to rigid diagnostic guidelines than others [67].

Thyroid stimulation hormone (TSH), CRP and cholesterol can be used as examples of how the extent of variation affects test results in practice when compared to guidelines for decisions.

The CV_I of TSH in adults is estimated in the literature between 16.2% and 29.3% [68, 69]. Although this is substantial, the NICE guidance on diagnosing primary hypothyroidism recommends a cut off value of over 10 mU/L for TSH [64]. However, the mean TSH level (in an American population) is only 1.5 mU/l [70]. Even with within-individual measurement variation of nearly a third, it would be rare to exceed the diagnostic threshold. It can be considered that the signal to noise ratio for this test is large and a single measurement is probably sufficient in most cases.

C-reactive protein (CRP) does not behave in this way, however. Like TSH, it also has a high within-individual measurement variation. A scoping review by the author showed a range of CV_I s for CRP of between 22 and 74.4% [46, 47]. CRP has a mean value in adults of 4.3mg/L [71], going up to 5.7mg/L in a population of 70 to

79 year olds [72]. For antibiotic prescribing, GPs often use a cut-off point of 10mg/L [73]. A mean close to the decision-making cut off point and a high variation therefore has the potential to lead to a high proportion of incorrect results. This is analogous to a medicine with a narrow therapeutic index and a variable bioavailability, which would lead in some individuals to a high risk of an adverse event.

Cholesterol has a range of CV_i s reported in the literature of 3-13.9% [74, 75]. Since cholesterol has a lower within-individual measurement variation than CRP, a single test for cholesterol is likely to be closer to the true patient mean than a single test of CRP.

It can therefore be seen that within-individual measurement variation has important implications for clinical decision-making. It is therefore vitally important to generate as much data as possible about the variation of clinical measurements, and to understand what problems this variation can cause for clinicians and patients when it comes to diagnosis, prognosis and monitoring decisions.

1.7.1 Existing variation data

Some measurands and areas of clinical practice have more data about the extent of variation than others. For example, blood pressure variation is well-studied [76]. The European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) also maintain a database of the variation of laboratory tests [77]. This critically appraises publicly available literature, extracts the important data and summarises the results. However, this only records the data from laboratory tests, and only includes the results of analyses performed in controlled conditions, mostly in healthy subjects, as opposed to using real-world data (that is data acquired in clinical contexts which are often less controlled than data acquired from experiments

performed under more highly controlled conditions) and patients with a variety of health conditions. By contrast there appears to be a smaller amount of data on real-world variation of laboratory tests, and a limited amount of data on the variation of function measurements and symptom scores/PROMs.

Rapid scoping reviews were carried out to assess the amount of published data on the within-individual variability for the following tests: Lipids; Thyroid Stimulating Hormone; HbA1c; Urinary albumin:creatinine ratio; C-reactive protein; Dementia symptom screening tests; Depression symptom screening tests; Spirometry measures. The methods and results for these are given in Appendix 10.1 and show that there is a highly limited amount of high quality published data (that is generalisable and with large sample sizes and large numbers of repeated measurements) on the within-individual variation of these tests. See 1.9 and table 1.3 for a summary of the results.

1.8 Sources of biological variability

There are many causes of non-random variability of a biological process or status within an individual over time. While much of the research into biological variation has taken place in the field of clinical chemistry [78], non-clinicopathological measurements also vary. Height has been shown to vary by up to an inch a day [79]. Resting heart rate varies within individuals [80]. Intraocular pressure (IOP) varies with body position and in a diurnal fashion, and even varies between eyes [81]. Spirometry values have also been shown to vary within individuals [82]. Within-individual measurement variation can also be important in imaging measures. Beat to beat variation has been demonstrated in echocardiographic measures of left

ventricular indices, for example [83], and gastric fluid volume and gastric emptying of water as measured by MRI show marked within-individual variation [84].

The following are the main causes of systematic variability within an individual over time.

1.8.1 Variability over lifespan

Many physiological values will change with age. Certain periods of life are associated with more rapid change, especially the neonatal period, childhood, puberty, menopause and old age [15]. Obvious examples are height and weight of children, but there is a myriad of other values that change over a lifespan. Other physical factors that vary with age include corneal thickness [85], bone mineral density [86], skin elasticity [87] and spinal flexibility [88]. Physiological and clinico-pathological factors that change with age include glycated haemoglobin HbA1c [14] and Forced Expiratory Volume in one second (FEV1) [14]. Testosterone declines at a rate of 1-3% per year in men [89] and in perimenopausal women, Alkaline Phosphatase (ALP) increases with age while thyroxine decreases with age [90]. Alzheimer's disease is a progressive disease and so measures of cognitive function such as the Mini Mental State Examination (MMSE) will be expected to show decline over time [91].

1.8.2 Circadian or diurnal rhythms

Circadian rhythms are related to time of day or sleep/wake cycles. Other daily changes in biological measurands may be related to posture (since plasma proteins are lower when lying down), food intake and activity levels [15]. Measurands measured in laboratories that vary over the course of a day in true circadian fashion

include thyrotropin, thyroxine and triiodothyronine [92], cortisol [93], neutrophils [94] and growth hormone [95].

1.8.3 Monthly cycles

It is well known that progesterone, oestradiol, luteinising hormone and follicular stimulating hormone vary in females on an approximately monthly cycle. Other hormones such as testosterone [96] and adiponectin [97] also vary in women with the menstrual cycle. Body temperature is a clinical characteristic that varies with the menstrual cycle [98] and depression measured by the Beck's depression score is an example of a symptom score that varies in some women over the course of a monthly cycle [99].

1.8.4 Seasonal cycles

A number of biological measurands vary with season. Since vitamin D is produced with the help of sunlight, it is not surprising that vitamin D levels have been shown to vary seasonally in various countries [100, 101]. Cholesterol also shows seasonality with 3-5% higher levels in the winter than the summer [102] and HbA1c, prolactin and free testosterone have shown cyclical seasonal variation in women [103]. Body weight and blood pressure have also been shown to vary seasonally [104, 105]. Depression symptom scores measured by the Patient Health Quality-9 (PHQ-9) test have been shown to vary in some populations with season [99].

1.8.5 Non-systematic or chance variability

Most measurands do not show significant systematic variation over time, and the major part of the variation is random fluctuation around the homeostatic set point [15]. Random biological variation has been described as the “noise” in our

physiological processes over time, occurring due to constant interactions with our internal physiology and external factors, such as weather, nutrition, exercise, stress and the large number of other environmental factors that surround us and impact us intermittently or continuously [1]. The body has a complex system of mechanisms to maintain homeostasis, so a perturbation away from the “ideal” set point of a body constituent (e.g. electrolytes, metabolites, hormones) will cause the body to institute mechanisms to move the level of the constituent back to the set point. For example, as mentioned in section 1.2.2, an increase in blood glucose levels will cause insulin secretion to decrease those levels, while a decrease in blood glucose levels will cause a reduction in insulin secretion and an increased secretion of glucogenic hormones to increase those levels, alongside other strategies to boost blood glucose such as inducing hunger. This variation around the set point is random.

Some biological values are more tightly controlled than others. For example, it is essential that calcium remains within a narrow range in the plasma, since hypercalcaemia causes disorders such as renal damage and soft tissue mineralisation and hypocalcaemia causes problems such as muscle weakness and seizures. Calcium has a coefficient of variation within individuals (CV_i) of 0.019, i.e. a very low within-individual variation. It is far less important to tightly control bilirubin however, and this consequently has a much higher within-individual variation with a CV_i of 0.26 [106].

1.9 Consistency of measures of variation

It has been stated that in general, estimates of within-individual biological variation are constant, regardless of size of study, country, time scale and methodology [15]. However, this has not been borne out by rapid scoping reviews by

the author of several tests performed commonly in practice. Table 1.3 shows the range in the CV_i reported in various studies of variation of laboratory tests, located by rapid scoping reviews.

Table 1.3 The intraindividual coefficients of variation (CVI) reported for selected laboratory tests.

Measurand	Range of all CVIs reported in studies (%)	References
Cholesterol	3.0-13.9	[74, 75]
HDL	2.75-11.00	[107, 108]
LDL	4.4-21.0	[108, 109]
Triglycerides	4.0-31.6	[110, 111]
TSH	16.2-32.8	[68, 112]
HbA1c	0.12-8.80	[113, 114]
ACR	11.3-42.0	[50, 115]
CRP	22.0-74.4	[46, 47]

Studies were diverse in terms of populations and disease status. HDL=High Density Lipoprotein; LDL=Low Density Lipoprotein; TSH=Thyroid Stimulating Hormone; HbA1c=glycated haemoglobin; ACR=urinary albumin:creatinine ratio; CRP=C-reactive protein.

Most studies reporting CV_I of a measurand do not give data on the uncertainty, e.g. the standard deviation or 95% confidence intervals. However, it has been shown that the degree of variation in blood pressure in individuals is not random, that is some individuals seem to have higher variation [116]. The degree of individual variation could be of relevance when monitoring individuals over a long term.

Furthermore, it is recently becoming recognised that the degree of within-individual variation in an individual can have prognostic importance. For example, the degree of variation in blood pressure has been shown to be associated with increased cardiovascular risk.

1.10 Measures of variation and their definitions

Variation can be described using the estimated mean and standard deviation or median and interquartile range. However, other summary estimates of variation can be more appropriate methods of describing the variation in certain circumstances. The most commonly used summaries that describe variation, with their definitions, are given in table 1.4.

Table 1.4 Measures of variation and their equations or definitions.

After Fraser [15], Sitch [98] and Po [99]

Summary	Definition
Variance (σ)	$\sum (X_i - \bar{X})^2$
Standard deviation (SD)	$\sqrt{\sigma}$
Coefficient of Variation (CV)	$\frac{SD}{mean}$
Index of Individuality (II)	$\frac{CV_I}{CV_G}$ or $\frac{\sqrt{CV_A^2 + CV_I^2}}{CV_G}$
Variation Independent of the Mean (VIM)	Variety of definitions
Reference Change Value (RCV)	$\sqrt{2Z} \sqrt{CV_A^2 + CV_I^2}$
Intraclass correlation coefficient (agreement)	$\frac{\sigma_G^2}{\sigma_A^2 + \sigma_I^2 + \sigma_G^2}$
Index of heterogeneity	$\frac{CV_{A+I}}{\sqrt{\frac{2}{n_2 - 1}}}$
Index of analytic error	$\frac{CV_A}{CV_I}$

CV_A =Analytical coefficient of variation; CV_I =within-individual variation; CV_G =between-individual variation.

Note that CV can be expressed as a percentage or a decimal fraction. For the systematic reviews and cohort studies (chapters two to nine) I have reported CV as a decimal fraction to avoid confusion with tests that are expressed as a percentage such as predicted FEV1 and FVC and HbA1c (NGSP units).

1.11 Research question, aims and hypotheses

How can clinical measurements be utilised better? How can we reduce over- and underdiagnosis and errors in monitoring and prognosis? How do we optimise the number of tests clinicians perform to increase accuracy while minimising wastage of

resources? Understanding within-individual measurement variation is one way to improve clinicians' use of clinical tests, with a resulting improved outcome for the patient, with reduced over- or undertreatment as well as more accurate diagnosis and prognostication.

In order to address these issues, seven measurands were selected for investigation in the primary care databases: HbA1c; CRP; FEV1 (Forced Expiratory Volume in one second); FVC (Forced Vital Capacity); PHQ-9 (Patient Health Questionnaire-9); HADS (Hospital Anxiety and Depression Score); BDI (Beck's Depression Inventory). The first two of these are laboratory tests, the second two are clinically assessed tests and the final three are symptom questionnaires (Patient Reported Outcome Measures or PROMs). They were selected because they are all commonly used tests in primary care, they are utilised to make diagnostic and treatment decisions, there is currently insufficient data on the within-individual measurement variation of these tests in primary care, and they are largely stable over long periods of time (but see chapters 3 and 7 for discussion of the effect of transient infections on CRP and chapters 4 and 8 for the effect of gradual long-term decline in lung function).

This thesis aims to add to the sum of knowledge in this area in the following ways:

- 1) Systematically review published evidence on within-individual measurement variation in seven measurands (HbA1c, CRP, Respiratory function tests FEV1 and FVC, Depression symptom measurement scores PHQ-9, HADS, BDI).
- 2) Assess within-individual measurement variation in these same seven measurands using a large dataset of medical records from primary care

- 3) Use simulated examples from the data as part of the analyses to illustrate the problems which within-individual measurement variation can cause.

1.12 Methodology overview

In order to answer the research questions regarding biological variation estimates, the IMRD and CPRD databases will be searched. The IQVIA Medical Research Database (IMRD) and the Clinical Practice Research Datalink (CPRD) are large, anonymised databases of electronic patient records (EPRs). The IMRD holds records from over 800 general practices in the UK. IMRD data has been shown to be generalisable to the general primary care population in terms of factors such as demographics, deprivation, condition prevalence and deaths [117]. The CPRD Aurum database captures routinely collected data from primary care practices in England and Northern Ireland [118]. It is broadly representative of the UK general population in terms of various diagnostic variables [119] [120] and has been shown to have high levels of correctness and moderate to high levels of completeness for selected conditions [121].

The IMRD and CPRD databases will be interrogated with the DexTER tool [122] to extract anonymised records of patients who have had repeated tests of the selected measurands. The data will be analysed in Stata to calculate a total measured within-individual coefficient of variation (CV_T).

The tests selected for this thesis to be studied are HbA1c, CRP, the spirometry tests FVC and FEV1, and the PHQ-9. These tests have been selected for matching the following criteria:

1. Tests commonly performed in general practice, and commonly repeated (as seen by interrogation of frequency of repeated tests in the IMRD database).

2. Represent a spread of both laboratory and non-laboratory clinical tests.

Investigating tests from a variety of clinical settings examines the extent of variation under different conditions. For laboratory tests, measurement is undertaken in standardised laboratory conditions, and since acceptable analytical precision is set at less than half of the biological variation, we would expect most of the observed variation to be biological [15]. Clinical measurements undertaken by medical staff have a degree of observer error due to variations in technique and skill, making reproducibility harder to quality assure. Patient reported outcome measures/symptom scores are self-reported, which introduces a different type of measurement variability.

3. Discussion with clinicians and examination of the literature shows that clinical decisions are made on the results of this test, rather than for example being one of a battery of tests on which a decision is made.
4. Scoping and systematic reviews (see Appendix 10.1, 20.1.1-20.1.8 have shown that the existing literature on the within-individual biological variation of these tests is limited.

1.13 Structure of thesis

In chapters two to five, the results of systematic reviews and meta-analyses of published studies describing variation are presented. Chapters two, three, four and five show the results of the systematic reviews of within-individual variation of HbA1c, CRP, spirometry scores and depression symptom measurement instruments respectively. In chapters six to nine, analyses of measurement variation in a large database of primary care records for the same measurands are presented and compared to the published data found in the systematic reviews. Chapters six, seven

and eight report the within-individual variation of HbA1c, CRP and spirometry measures respectively as extracted from the IMRD database. Chapter nine shows the within-individual variation of depression symptom measurement instruments extracted from the CPRD database.

In chapter 10 the findings of these studies are summarised and the implications of the findings for clinical practice and clinical guidelines creation are discussed. The strengths and limitations of the research are also discussed and recommendations for further research are made.

In the next four chapters in this thesis, I report the systematic reviews that show the previously published data on within-individual measured variation in the measurands of interest, starting with one that has some previously published data and systematic reviews, the marker of average glucose levels, HbA1c.

2 Chapter two: Within-individual variation of HbA1c: A systematic review and meta-analysis

A version of this chapter has been previously published as:

Gough A, Sitch A, Ferris E, Marshall T. Within-subject variation of HbA1c: A systematic review and meta-analysis. PLoS One. 2023 Aug 2;18(8):e0289085. doi: 10.1371/journal.pone.0289085.

2.1 Abstract

Background

Glycosylated haemoglobin (HbA1c) measurement is used to diagnose and to guide treatment of diabetes mellitus. Within-individual variation in measured HbA1c affects its clinical utility and interpretation, but no comprehensive systematic review has described within-individual variation.

Methods

A systematic review and meta-analysis was performed of within-individual variation of HbA1c. Multiple databases were searched from inception to November 2022 for follow-up studies of any design in adults or children, with repeated measures of HbA1c or glycosylated haemoglobin. Title and abstract screening was performed in duplicate, full text screening and data extraction by one reviewer and verified by a second. Risk of bias of included studies was assessed using a modified consensus-based standards for the selection of health measurement Instruments (COSMIN) tool. Intraclass correlation coefficient (ICC) results were pooled with a

meta-analysis and coefficient of variation (CV) results were described by median and range.

Results

Of 2,675 studies identified, 111 met the inclusion criteria. Twenty-five studies reported variation data in healthy patients, 19 in patients with type 1 diabetes and 59 in patients with type 2 diabetes. Median within-individual coefficient of variation (CV) of HbA1c was 0.070 (IQR 0.039 to 0.090). For healthy subjects the median CV for HbA1c was 0.017 (IQR 0.013 to 0.022), for patients with type 1 diabetes 0.084 (IQR 0.067 to 0.090) and for type 2 diabetes 0.083 (IQR 0.060 to 0.100). CV increased with mean population HbA1c.

Limitations

Assessment of variation was not the main aim of many of the included studies and some relevant studies may have been missed. Many included studies had few participants or few repeated measurements.

Conclusions

Within-individual variation of HbA1c is higher for patients with than without diabetes and increases with mean population HbA1c. This may confound observed relationships between HbA1c variation and health outcomes. Because of its importance in clinical decision-making there is a need for better estimates and understanding of factors associated with HbA1c variation.

2.2 Introduction

Haemoglobin A1c (HbA1c) is produced by non-enzymatic glycation of haemoglobin. It provides an estimate of glycaemia (mean glucose levels) over the preceding one to three months, proportionately weighted to more recent periods [123]. It is therefore used to diagnose diabetes mellitus and to monitor patients with diabetes. The American Diabetes Association (ADA) recommends a cut-off point of more than or equal to 6.5% (48mmol/mol) HbA1c level for a diagnosis of diabetes mellitus and 5.7–6.4% (39–47 mmol/mol) for a diagnosis of prediabetes [124]. The Expert International Committee (EIC) recommends a cut off of 6% (42mmol/mol) for prediabetes and the World Health Organisation (WHO) does not recommend HbA1c use for diagnosis of prediabetes [125]. It is recommended that HbA1c is measured every three to six months in newly diagnosed patients with type 2 diabetes mellitus until stabilised, and thereafter every six months [126].

HbA1c levels can be reported using either International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), expressed in mmol/mol, or National Glycohemoglobin Standardization Program (NGSP) reference systems, expressed as % [127]. Although a conversion formula between the two methods is available, they may not be directly interchangeable [128]. Further, within-individual variation in NGSP units is reported to be lower than in IFCC units [129]. Although the IFCC is considered to be the higher standard test, many countries continue to use NGSP units [130].

Reference ranges for diagnostic tests are usually set by comparison to a reference population, which allows assessment of between-individual variation. However, biological processes and values also vary over time. This variation can be

systematic and predictable, such as seasonal variation [131], or may be due to chance. This longitudinal within-individual variation is known as biological variation. Variation can also be introduced into a measurement from pre-analytical factors such as stress, exercise and food intake prior to the laboratory measurement and posture during the sampling procedure. Imperfect accuracy and precision of laboratory measurement mean that if a measurement is repeated on the same sample, there will be a range of results around the true value (analytical variation).

The variation in measured HbA1c results encountered in real-world clinical practice is because of a combination of biological within-individual variation, pre-analytical and analytical variation. Biological variation is usually (but not always) the greatest component of total variation, since clinical laboratories usually set acceptable analytical variation at <0.5% of biological variation [15].

The greater the within-individual variation of a measurand, the lower the probability that a single measurement is an accurate reflection of the true mean of the measurand in that individual. This can lead to errors in diagnosis, prognosis and treatment. Specifically for HbA1c, an inaccurate result may lead to an incorrect, missed or delayed diagnosis of diabetes mellitus, causing inappropriate commencement of treatment or an inappropriate delay in commencing treatment, or an incorrect estimation of glycaemic control in patients with diabetes, leading to an inappropriate increase in treatment intensity or a delay in intensifying treatment.

A scoping review searching for studies on within individual variation of HbA1c (see Appendix 10,1, 20.1.5) showed a small number of studies on the subject of within individual variation of this measurand. A search of Pubmed, Epistemonikos and the Cochrane database revealed no previous systematic reviews of within-individual

variation of spirometry measures showed that three previous systematic reviews on HbA1c variation have been published. Gonzalez-Lao et al [132] reported the results of a systematic review and meta-analysis of 17 studies. Based on three studies Gonzalez-Lao et al gave a pooled estimate of within-individual coefficient of variation (CV_i) for HbA1c in healthy adults of 0.013 (0.012-0.025) for results reported in IFCC units (mmol/mol) and based on four studies, a CV of 0.013 (0.012-0.021) for results reported in non-IFCC units (%). CV was slightly higher in patients with diabetes mellitus than subjects without diabetes. The European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) database reports a CV_i of 0.016 (0.013-0.024) for studies reporting IFCC units based on four studies and 0.012 (0.003-0.019) for studies reporting NGSP units (based on 7 studies) [77]. Braga et al performed a systematic review without meta-analysis of eight studies, with CV_is ranging from <0.007 to 0.098 [133].

There is a need for an up to date and more comprehensive systematic review of within-individual variation of HbA1c. Note that this pragmatic review allows the inclusion of real-world data to improve the clinical applicability of the results.

2.3 Aim of review

To describe and evaluate the current literature on within-individual variation of measured HbA1c and from this to estimate variation in people with and without diabetes mellitus.

2.4 Methods

PRISMA and PRISMA-S guidelines were followed in the reporting of this review [134].

Searches were devised to identify cohort studies, clinical trials or any studies in which an HbA1c or glycosylated haemoglobin measurement was performed more than once in the same individual. See Appendix 2.1 for search strategies. The searches for this study were combined with the study into C-reactive protein (CRP) variation, and then CRP studies were excluded at the full text screening stage.

The international prospective register of systematic reviews, PROSPERO, was checked for ongoing reviews, and the protocol was registered with PROSPERO.

CRD42020202387. The review protocol is available at

https://www.crd.york.ac.uk/prospERO/display_record.php?ID=CRD42020202387

On 5th August 2020, database searches were carried out. Medline, Embase, Cochrane Central, Epistemonikos and Open Grey were searched from inception to 5th August 2020. An update to the search was performed to include studies published up to November 2022 (full details in Appendix 2.2). Subject experts were contacted for suggestions for further studies. Search terms were adapted for each database searched. The references of included studies were checked by hand for further relevant studies. Endnote reference management software was used to collate studies revealed by the search.

Studies were included if primary research data on the variation of at least two measurements of HbA1c or glycosylated haemoglobin within the same subject was recorded. Studies could be of any design. The population included adults and children, healthy or with any disease condition, in any setting. The outcome was

variation which could be reported as coefficient of variation (CV), standard deviation (SD), variation independent of the mean (VIM), index of individuality (II), Reference Change Value (RCV), index of heterogeneity, validity coefficient (VC), ICC agreement, ICC consistency, Cronbach's alpha or Cohen's kappa. There was no restriction on time of publication, language of publication, population, setting or sample size. Studies were excluded if participants were not in a steady state (measurements were before and after an intervention or had an acute or rapidly changing illness) or data were secondary (systematic and narrative reviews).

Studies were grouped for synthesis based on variation measure reported e.g. ICC, CV. Since this study aimed to capture variation data wherever it was published, there was no limit on publication date, language or study design.

Titles and abstracts were screened independently by two reviewers (AG and EF) in Abstrackr systematic review software [135], and those identified of interest underwent full text searches. Full texts were screened by AG and exclusions confirmed by EF. Differences were resolved by discussion. Foreign language studies were translated by Google translate software. Data were extracted into Excel from the full text of the study studies where possible, and from abstracts where full texts were not available, by a single reviewer (AG). All data extraction was verified independently by a second reviewer (EF). Where the same study was reported in multiple studies, the full text paper was preferred over an abstract, English language preferred over non-English, and the earliest English version over the later if there was more than one.

Table 2.1 lists the outcome and other main variables extracted. Variables selected for extraction and for subgroup analysis were those hypothesised to most

likely affect variability. All eligible outcomes were included where more than one outcome was reported within a paper. The primary outcome measure was variation of repeated measurements within the same individual. Where multiple measures of variation were given in a single study, the primary population only was analysed for the main meta-analysis. The primary population was the full study population (as opposed to subgroups), and if the full study population was not given, the primary outcome was identified using the following hierarchy: 1. Healthy population; 2. Most stable population (i.e. subjectively judged to be in the most steady-state such as disease course or treatment); 3. First outcome listed in the paper.

Table 2.1 Variables extracted from studies

Variable	Definition and rules
Study design	Cohort, RCT etc
Number of subjects	Number of subjects used to calculate variation measure
Age	Average age of subjects used to calculate variation measure. If the subjects used to calculate variation measure is a subset of the study population and only the age is given for the whole population, the age of the whole population is used.
Sex	Percentage of subjects used to calculate variation measure who were male. If the subjects used to calculate variation measure is a subset of the study population and only the sex is given for the whole population, the sex of the whole population is used.
Ethnicity	Ethnicity recorded yes/no? If country of origin of subjects only is recorded, this is counted as no.
Setting	Primary care/community versus secondary/tertiary/laboratory setting
Health status	Healthy, type I diabetes mellitus, type II diabetes mellitus or other
Number of measurements	Number of repeated HbA1c measurements in the same individual
Maximum time interval	Maximum time interval between consecutive measurements
Minimum time interval	Minimum time interval between consecutive measurements
Method of variation calculation	Is the variation calculation described?
Unit of measurement	Mmol/mol or %?
Variation described as total or component parts?	Is variation calculation explicitly described as total variation, within-individual variation, or uncertain?
CV	Coefficient of variation of repeated measures within an individual
SD	Standard deviation of repeated measures within an individual
ICC	Intraclass correlation of repeated measures within an individual

Where information was missing or unclear, the information was not extracted with the following exception: where patient characteristic data was reported only for a

whole study population, but variation data was only present for a sub-population, the patient characteristic data for the whole population was extracted and used for subgroup analyses.

There is no risk of bias tool currently in existence that is ideal for this review. A bespoke risk of bias tool was developed and included in the original protocol recorded in Prospero (Appendix 2.3b), but it was ultimately decided preferable to use a validated risk of bias tool. A risk of bias tool was used that was adapted from the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) risk of bias tool for test reliability [136] (Appendix 2.3a). This tool enabled some estimate of risk of bias for each paper, although the questions on blinding were problematic, given that none of the studies included stated that the investigators performed the measurements or assigned scores without knowledge of previous scores by the individual. The risk of bias scores are therefore presented in this study with the two questions on blinding excluded. However, there is no reason to think a lack of blinding would alter the measurements in this study, given the tests are not being performed to demonstrate a difference in a clinical trial.

The main outcome measure of this review is coefficient of variation of repeated measurements within the same individual. All studies were considered for synthesis provided they had a measure of variation that had a sufficient number of studies reported so they could be pooled. Coefficients of variation were converted from % to a decimal fraction.

Stata SE 17 was used to perform statistical calculations and generate forest plots. ICCs were transformed using Fisher's Z transformation [137]. Z scores were then back-transformed (Appendix 2.6) and 95% confidence intervals of the ICC were

calculated. Coefficient of variations were considered for meta-analysis, but no well-described and validated methods currently exist for meta-analysis of coefficient of variations, so the analysis of these is descriptive. Descriptive subgroup analyses were performed based on unit of measurement (IFCC or NGSP), health status, setting, number of measurements and whether short or long-term variation was measured. Sensitivity analyses were based on number of subjects and risk of bias.

2.5 Results

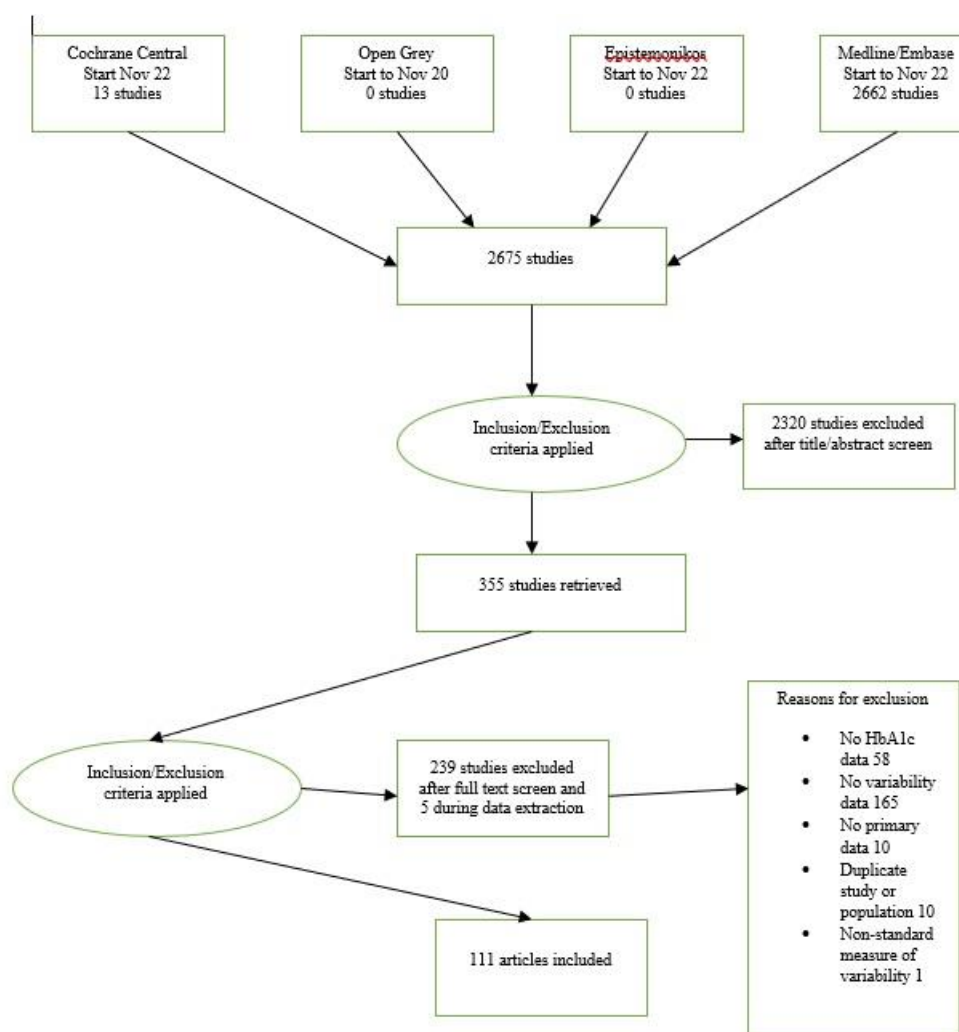


Figure 2.1 PRISMA flow diagram describing selection of studies.

After database searches and hand searching reference lists of included studies, 2,675 non-duplicate studies were identified of which 2,320 were excluded after title and abstract screening leaving 355 studies. After full text screening, 244 were excluded, and a further five during data extraction (see fig. 2.1). Most full text exclusions were because no recognised measure of variation was reported.

One hundred and eleven studies met the inclusion criteria (31.3% of full text studies screened). Sample sizes ranged from 4 to 91,866 individuals, with a median of 378. One hundred-and-five studies (95%) reported longer term variation (arbitrarily defined for this study as seven or more days between measurements), two reported short-term variation (less than seven days between measurements), one reported both long- and short-term, and three did not report measurement intervals. One hundred and six studies (95.5% of included studies) were of a cohort design, three (2.7%) used data from the placebo arm of a randomised controlled trial and two (1.8%) were case-control studies. Twenty studies (18.0%) were carried out in a primary care or community setting and 73 (65.8%) studies were carried out in secondary or tertiary setting such as universities, hospitals and laboratories. In eighteen (16.2%) studies the setting was not reported.

Study populations were diverse in terms of age, gender and health status. Ethnicity was recorded in 34 (30.6%) studies. Included studies and full study characteristics are presented in S2.4. With questions four and five of the risk of bias scoring excluded, 21 (18.9%) studies scored the best risk of bias score of one, 53 (47.7%) studies scored a risk of bias of two, and 37 (33.3%) studies scored a risk of

bias of three. Risk of bias scores are presented in Appendix 2.5. Table 2.2 summarises the study characteristics of the included studies.

Table 2.2 Summary statistics of study characteristics of included studies

Study characteristic	Studies reporting/111	Median	Lower quartile	Upper quartile	Minimum	Maximum
Number of subjects	110 (99%)	473	59	2103	4	91866
Age (years)	84 (76%)	59.05	44.00	64.79	8.77	76.00
% male	88 (79%)	52.34	46.49	58.11	0	98.00
Number of measurements	98 (88%)	5.0	4.0	10.4	2.0	36.8

Included studies consisted of ninety-four full texts and seventeen abstracts. The average age of subjects in the primary study populations ranged from 8.8 to 76 years, with a median of 59.1 years. The percentage of subjects who were male ranged from 0% to 98% with a median of 52.3%. Ninety-three studies reported the setting (e.g. primary/community care, secondary/tertiary care, laboratory university setting). Twenty of the 93 (21.5%) studies that recorded setting were conducted in primary care or the community and 73/93 (78.5%) in secondary or tertiary care or in a laboratory/university situation. Health status of individuals was recorded in 110/111 studies of which 25 (22.5%) were healthy, 59 (53.2%) had type 2 diabetes, 19 (17.1%) had type 1 diabetes, 4 (3.6%) had both type 1 diabetes and type 2 diabetes, and 3 (2.7%) had other diseases.

Aims of the studies included explicitly measuring biological variation, investigating associations between biological variation and outcomes such as mortality and other events (e.g. albuminuria, nephropathy), identifying risk factors associated with increased biological variability (such as living alone) and calculating reference change values.

The average number of measurements ranged from 2 to 36.8 (i.e. the study with the largest average number of measurements had an average of 36.8 measurements per person) with a median of 5. Eighty-four (75.7%) studies used

NGSP (%) units to calculate the variation and four (3.6%) used IFCC (mmol/mol); 23 (20.7%) reported in both units or did not clearly specify the units.

Eighty-three (74.8%) studies reported total variation, 15 studies (13.5%) reported individual variation, and 1 (0.9%) study reported both. For 12 (10.8%) studies the type of variation reported was unclear. The CV of the HbA1c primary population ranged from 0.003 to 0.179, with a median of 0.070 (IQR 0.051). The within-individual standard deviation of the HbA1c of the primary population where the units were NGSP (%) ranged from 0.08 to 1.47 with a median of 0.62 (IQR 0.37). Only five studies reported ICC, with a median of 0.87 and a range of 0.59 to 0.97.

One paper [138] appears to report an incorrect value for the CV since the value appears to be implausibly low compared to other reports of this result and is also inconsistent with the other variation measures reported in the paper. The value for CV from this paper used in this review was therefore calculated from the reported SD and mean. Two studies did not report a CV, ICC or SD. Segar [139] reported variation as average successive variability (average absolute difference between successive values), with a result of a mean of $0.6 \pm 0.3\%$ (mean $\pm$ SD). Sugawara [140] used an adjusted standard deviation to account for different numbers of HbA1c measurements, reporting an adjusted SD result of $0.79 \pm 0.6\%$ (mean $\pm$ SD). Three studies [141-143] reported variation measurements for units of both mmol/mol and % and the % unit is included in this review.

The median within-individual coefficient of variation (CV) of HbA1c was 0.070 (IQR 0.039 to 0.090). Table 2.3 summarises the results of variation measures of the primary study population.

Table 2.3 Summary of results of variation measures of primary study population

Study result	Number of studies reporting result (out of 111)	Median value of result	IQR	Minimum	Maximum
CV	89	0.0698	0.0512	0.0029	0.1790
ICC	6	0.81	0.19	0.59	0.97
SD (% units)	35	0.62	0.37	0.08	1.47

With units reported as %, CV in healthy subjects (median 0.017; IQ range: 0.013 to 0.022) was lower than in patients with type 2 diabetes (0.083; IQ range: 0.06 to 0.10), type 1 diabetes (0.084; IQ range 0.067 to 0.89) or other health conditions (0.079; IQ range 0.048 to 0.11) (Fig 2.2 and table 2.4). The number of studies reporting mmol/mol units was too small for meaningful conclusions(Fig 2.3).

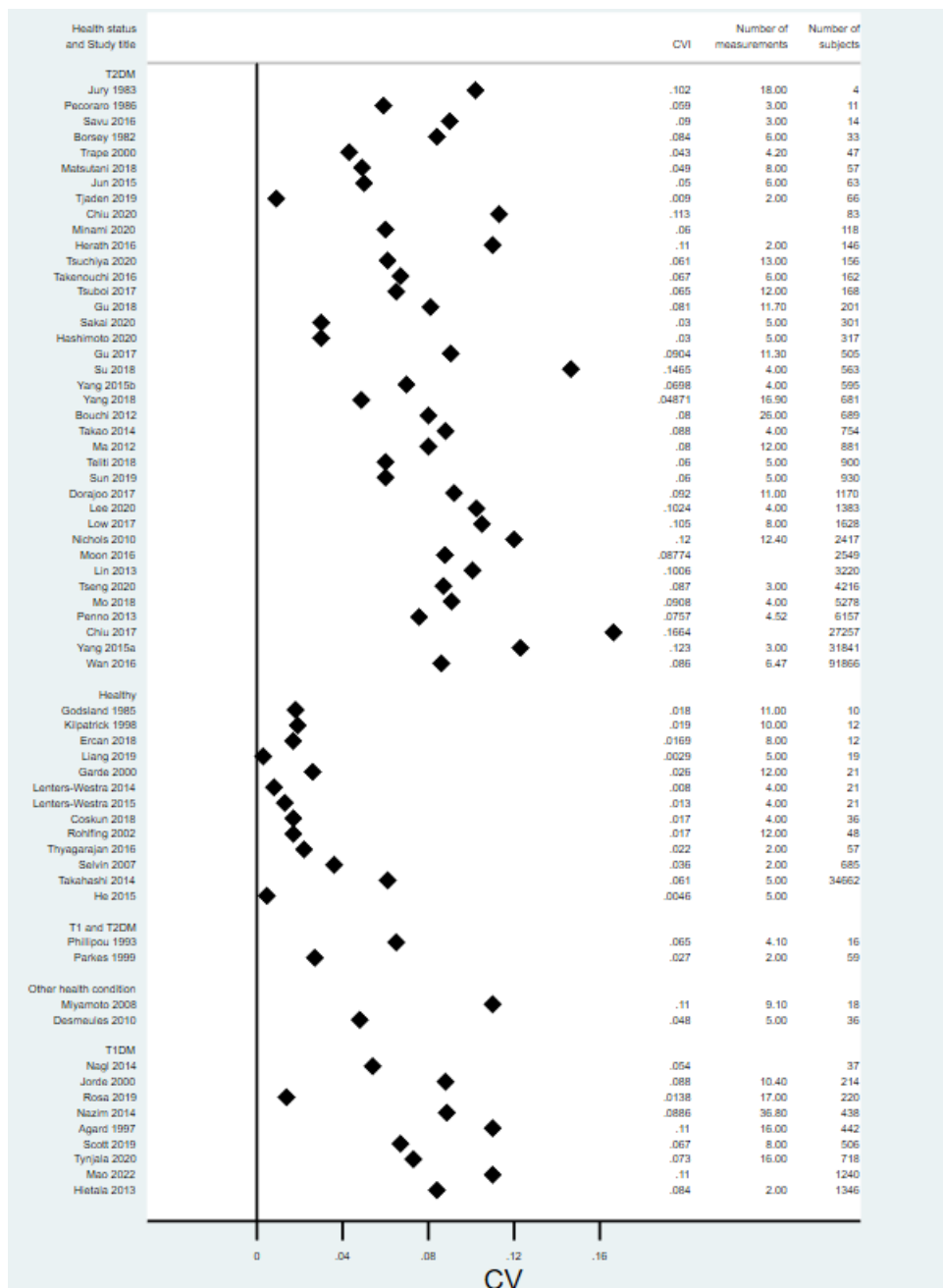


Figure 2.2 Forest plot of CV for studies reporting in units of %, split by health status.

T1DM=type 1 diabetes; T2DM=type 2 diabetes

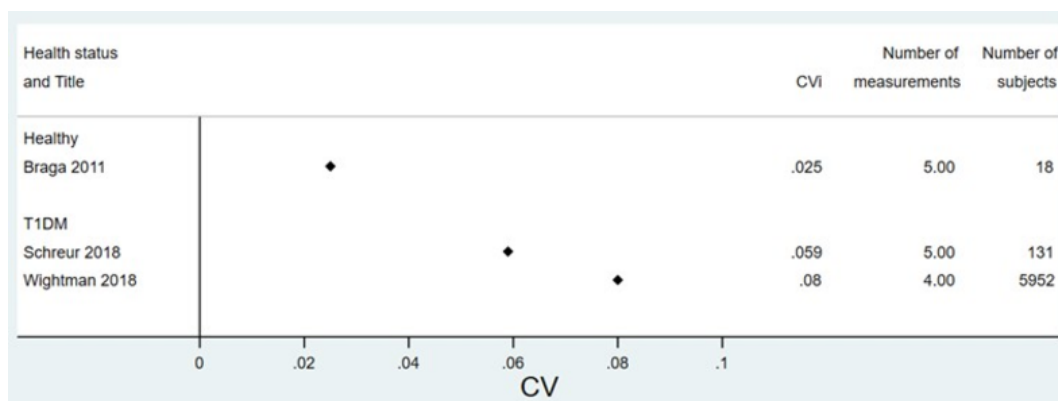


Figure 2.3 Forest plot of CV for studies reporting in units of mmol/mol, split by health status

Table 2.4 Summary statistics of reported CV reporting in units of %, split by health status

Health status	Observations	Median CV	IQR	Min	Max
Healthy	13	0.017	0.009	0.003	0.061
Type 2 diabetes	38	0.083	0.040	0.009	0.166
Type 1 diabetes	9	0.084	0.022	0.014	0.110
Types 1 and 2 diabetes	2	0.046	0.038	0.027	0.065
Other	2	0.079	0.062	0.048	0.110

Figs 2.4 and 2.5 show results for studies measuring variation as SD or ICC. Forest plots for CV for studies reporting units of %, stratified by risk of bias, setting, short- or long-term measurements and whether CV_I or CV_T was reported or whether this was uncertain can be found in Appendix 2.7 (figs 12.1 to 12.4). Subjectively, lower risk of bias was associated with lower variability (fig 12.1). There was no apparent difference between the variation in primary care and other settings (Fig 12.2). Too few short-term studies were included to be able to draw any conclusions about effect of length of time between measurements on variations (fig 12.3). Studies explicitly measuring within-individual rather than total variation of HbA1c had

lower CVs (fig 12.4) which is expected since they only report one component of the variation, and also studies calculating within-individual variation are likely to have been performed under more controlled conditions.

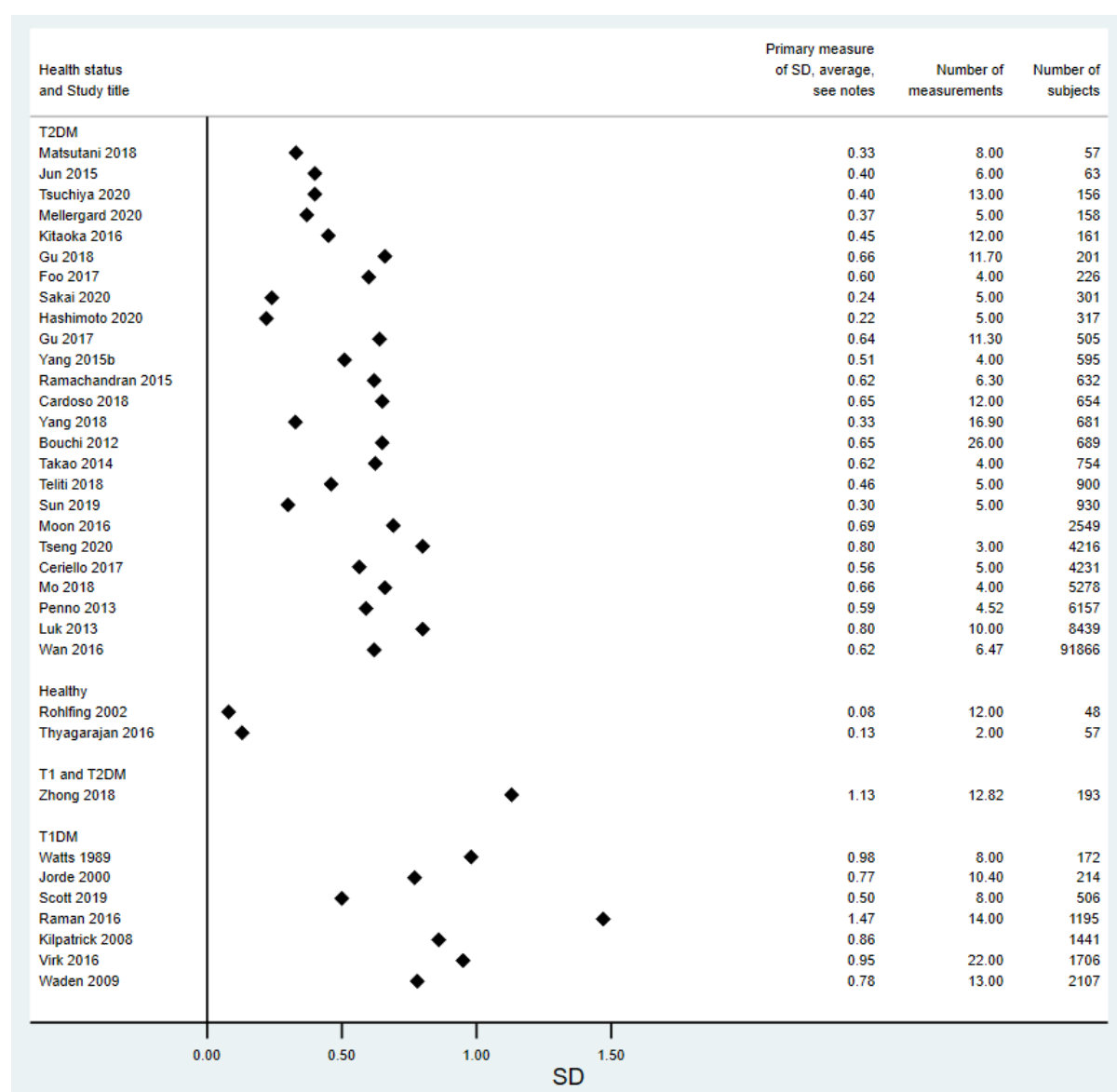


Figure 2.4 Forest plot of SD for studies reporting in units of %, split by health status.

T2DM = type 2 diabetes mellitus, T1DM = type 1 diabetes mellitus

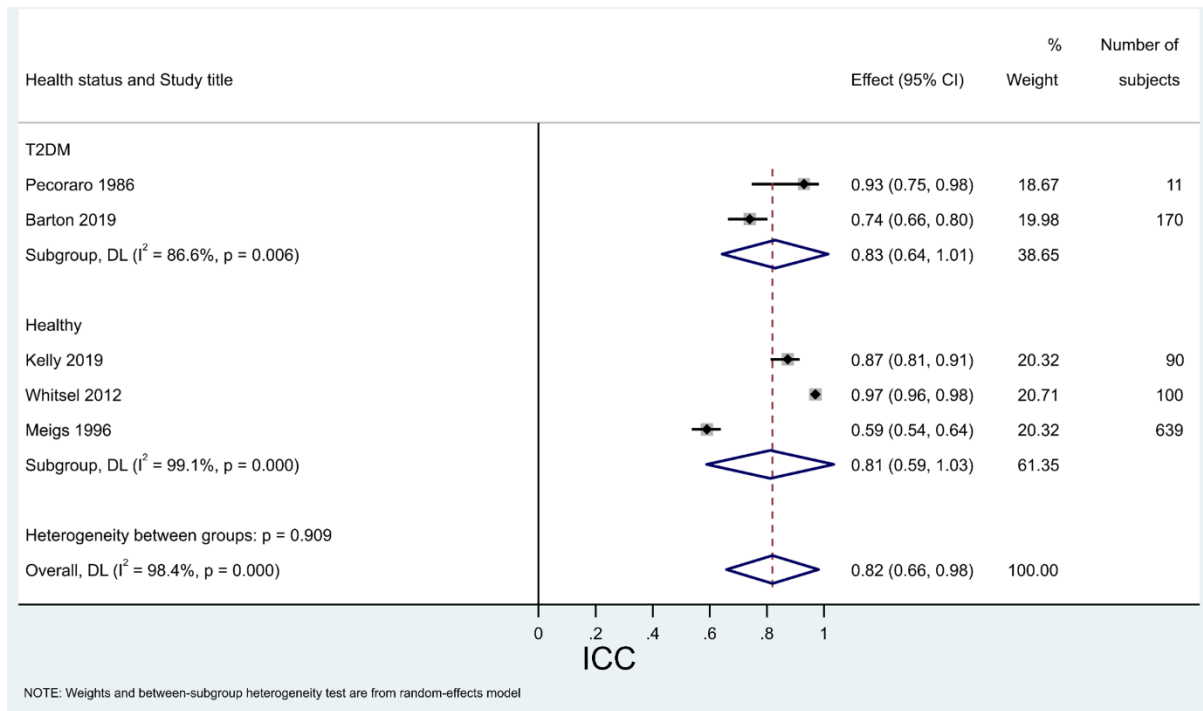


Figure 2.5 Forest plot with meta-analysis of ICC for studies reporting units of %, split by health status.

T2DM = type 2 diabetes mellitus

Because the variation was higher in patients with diabetes, for all studies where it was reported I plotted mean population HbA1c against CV (Fig 2.6). This demonstrated a moderate positive correlation ($R^2 = 0.48$), showing that CV increases with mean HbA1c.

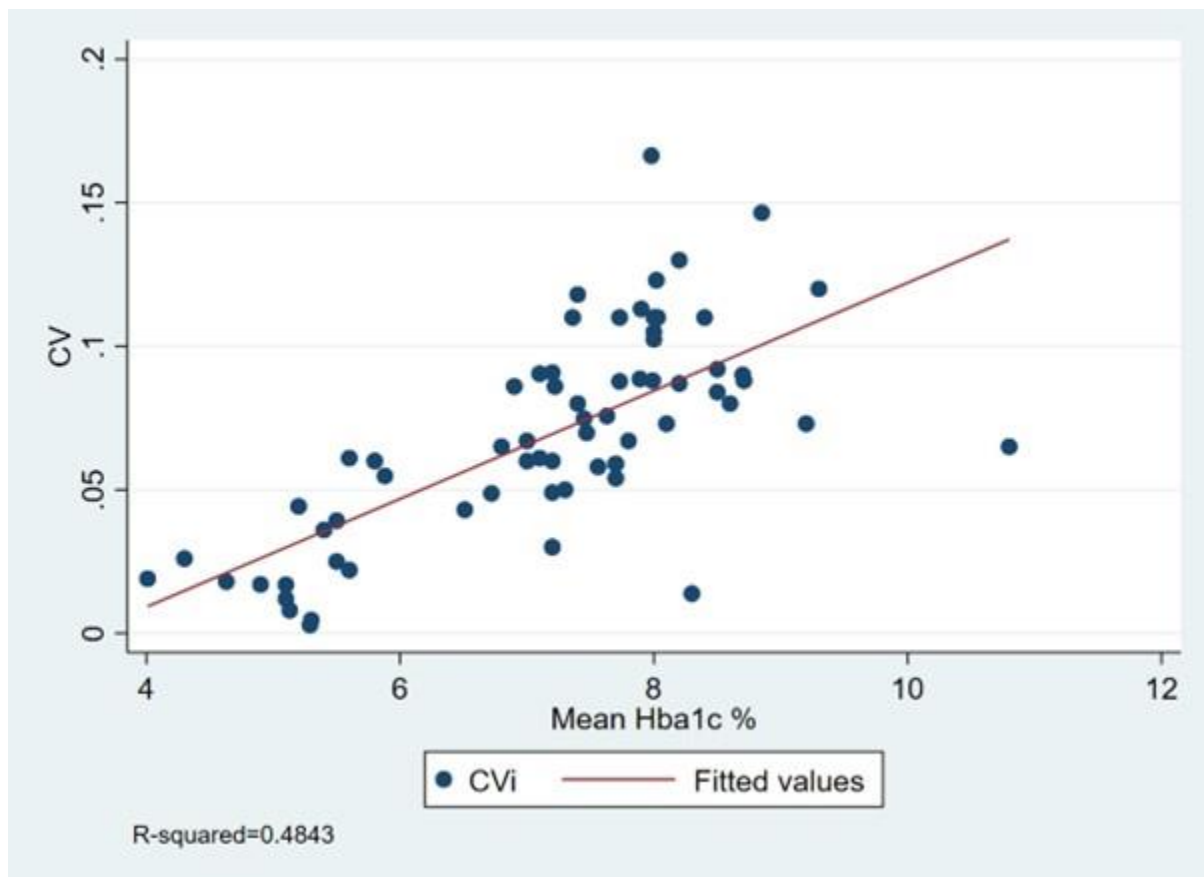


Figure 2.6 Scatter plot of mean population HbA1c against CV, with line of best fit.

2.6 Discussion

This review is the most comprehensive systematic review and meta-analysis of within-individual HbA1c variation. Of the three previous systematic reviews, one did not perform a meta-analysis and two only performed meta-analyses on healthy populations. The largest included only 17 studies in total, three in the meta-analysis of variation measured using of mmol/mol and four in the meta-analysis of % units [132]. This review included 111 studies. Previous systematic reviews on HbA1c variation were designed to estimate a coefficient of variation under ideal conditions in order to inform analytical performance specifications, reference change values and population-based reference intervals. By contrast the current study was designed to be as broad as possible, increasing generalisability by including HbA1c variability in a variety of settings and health conditions.

There is no well-described, validated method for meta-analysis of CVs therefore it was not undertaken in this study.

The median estimates of CV (NGSP %) were 0.017 (IQ range: 0.013 to 0.022) in healthy individuals, 0.083 (IQ range: 0.060 to 0.100) in type 2 diabetes and 0.084 (IQ range 0.067 to 0.890) in type 1 diabetes. These are similar to previous reviews of healthy patients which report a median CV of 0.013 (95% CI 0.012-0.025) for IFCC (mmol/mol) units and 0.013 (0.12-0.21) for NGSP(%) [132]; 0.016 (95%CI 0.013-0.024) for IFCC(mmol/mol) and NGSP (%) 0.011 (0.002-0.019) for % [77]. This is consistent with previous research suggesting that patients with diabetes have a higher CV of HbA1c than healthy patients [144] [132].

I demonstrated a correlation between mean population HbA1c and within-individual variation measured by CV. This suggests individuals with prediabetes

have a higher HbA1c variability than the healthy individuals, and poorly-controlled patients with diabetes have a higher variability than well-controlled patients with diabetes. This may be due to different levels of insulin resistance and/or variation in caloric intake. It also means clinical outcomes which correlate with mean HbA1c will also correlate with CV. Researchers investigating variability as a predictor of outcomes should consider using a measure of variability which is independent of the mean.

Although laboratory tests such as HbA1c are usually performed under the same laboratory conditions regardless of setting, there are likely to be significant differences between control of pre-analytical factors and populations in studies performed in primary care and those performed in secondary or tertiary settings or in research settings where the express aim is calculation of biological variation. For example, pre-analytical conditions such as patient stress, recent diet and exercise, posture and sample storage are all likely to differ between settings. Further the population in primary care is less likely to be acutely or severely ill than those in referral settings. However, there are too few studies in primary care in this review to draw meaningful conclusions about the differences between settings.

The variability of HbA1c has implications for clinical decision-making, with potential for misdiagnosis if a measured value is obtained that is not consistent with the patient's true mean. General practitioners' understanding of this issue derived from clinical experience may explain why they often do not commence treatment with oral anti-diabetic drugs until they document a HbA1c reading well above the cut-off recommended by National Institute for Healthcare and excellence (NICE) guidelines [145]. Of even more importance are the implications for monitoring and treatment decisions, since the variability is higher in these patients than healthy individuals.

Metformin monotherapy lowers HbA1c by 1.1% NGSP units (12mmol/mol IFCC units on average) [146] which is within the usual range of within-individual variation. This is consistent with the finding that 15-18% of patients that commence treatment with oral anti-diabetic drugs demonstrate an increase in measured HbA1c – this can be explained by within-individual variation.

Another concern is that higher HbA1c variability is associated with an increased all-cause mortality due to a variety of mechanisms such as endothelial dysfunction, increased oxidative stress and increased release of cytokines [147], demonstrating the importance of reducing HbA1c variability by improving control of diabetes. Moreover, a strict glycaemic control may play a cardioprotective effect [148].

Two obvious outliers were found in this review, one of which showed a CV of 0.59 [138]. This appeared to be an error and the CV was recalculated for this review from data available in the study to give a result of 0.086. The other study showed a CV of 0.0029 [113], although there is an error in the abstract where a different CV result is given, which appears to be the result for males only. The authors acknowledge their result is lower than previously published estimates of HbA1c CV and suggest this may be due to strict control of pre-analytic factors and Chinese ethnicity of the subjects. This latter seems unlikely since four other studies with Chinese subjects reported a CV ranging from 0.075 to 0.17 [138, 149-151]. Nevertheless, it has been shown that ethnicity influences HbA1c levels [152]. This does illustrate the importance of taking ethnicity into account when studying variability, and only 34/111 studies in this review explicitly reported the ethnicity of their subjects, although it is not always clear whether nationality or ethnicity is being referred to. Future research in HbA1c variability should include the ethnicity of subjects.

One weakness of this systematic review is that because variability was not the primary outcome of the studies some relevant studies may not have been identified. There are also a number of weaknesses with some of the studies in this review. Some included studies had low numbers of subjects and/or low numbers of repeat measurements. It was not always clear whether studies were reporting total variability (i.e. $CV_A + CV_I$) or within-individual variability (CV_I). Fig 12.4 in Appendix 2.7 illustrates the CVs of studies split by whether they reported CV_I , CV_T , or whether this was uncertain. Subjectively there seemed to be minimal difference between the groups.

Note that some studies may not have clearly distinguished between a material change in the severity of the disease in populations that have the disease (e.g. an increase in mean value) and an increased level of variation due to the disease (the two aspects of pathological variation), which could increase the measured variability and influence the association of increased variability with presence of disease/increased disease severity.

Note also that some of the underlying conditions for which the tests examined in this thesis are utilised are much more variable than diabetes.

HbA1c may not be an ideal measure of glycaemic control because it does not capture hypoglycaemic events. In the future it may be replaced by flash glucose monitoring for some patients, but at present it is the main test used to assess glycaemic control and is likely to remain so for most patients.

Twenty-three studies did not report the method with which they calculated variability. This is particularly important with ICCs since there are at least ten different ways of calculating ICC which can all give different results [153]. Most

studies reported results in units of % (NGSP method) despite mmol/mol (IFCC method) being recommended. In some studies, it was unclear whether results were expressed using NGSP or IFCC methods, which is important since these methods can give different results.

A bespoke risk of bias tool was developed and included in the original protocol recorded in Prospero (Appendix 2.3b), but it was ultimately decided preferable to use a validated risk of bias tool, and the COSMIN tool was selected (Appendix 2.3a) since it was validated and was more applicable to a wider range of tests than other tools considered. The bespoke tool scored bias in five domains: patient selection; index tests; flow and timing; statistical analysis; conflicts of interest, with the risk in each domain ranked as low, high or unclear. The COSMIN tool consisted of two sections. Part A collected information on elements of the research question such as the method of analysis and the patient population. Part B rated the bias on a scale of 1 to 4 (very good to inadequate), with questions including “Were the patients stable in the time between the repeated measurements?”, and “Were there any other important flaws in the design or statistical methods of the study?”. The COSMIN tool has some limitations, particularly when it comes to the questions regarding blinding. Generally, in studies into biological variation the investigators are not blinded to previous measurements. These questions were therefore omitted when a study’s risk of bias score was assigned. This is unlikely to have affected findings as HbA1c is measured objectively in a laboratory. However, knowledge of previous test scores might influence the decision to perform the test in the first place, which could lead to confounding by indication. A third risk of bias tool. The BIVAC instrument was considered [154]. This has been designed to assess risk of bias in studies of biological variation, but only considers the CV as the metric of variation,

which would exclude other measures of variation. Further, it is only suitable for laboratory tests and so I chose to utilise a risk of bias tool that was relevant to all four systematic reviews.

Risk of bias due to missing publications was deemed to be low, since in most cases the outcome measure of this review was not the primary outcome of the individual studies, and therefore likelihood of publication is unlikely to be related to variation outcomes. Some studies with repeated measures do not report variation data, but this is unlikely to be related to variation outcomes.

As some publications included multiple populations, with information being given only on the subgroups and not the total, arbitrary decisions had to be made as to which populations to include in the primary analysis. However, the rules to decide which group to include were applied systematically, using the following hierarchy: 1. Healthiest population; 2. Most stable population; 3. First listed population. Some studies reported demographic data for the whole population but not for the subgroup for which variation data was reported. In these cases, the demographic data for the whole study population was used and was assumed to be close to the variation subgroup. Although a larger number of studies has been included than in previous reports, the quality of the studies and the other limitations of the review mean that an accurate real-world estimate of within-individual variation of HbA1c remains lacking. Undertaking further studies with larger numbers of subjects with a variety of health conditions would help to understand the extent to which HbA1c varies within an individual over time in a clinical setting. This would help inform clinicians when making diagnostic and treatment decisions, and the authors of clinical guidelines when making recommendations on diagnosis and treatment.

2.7 Conclusion

This comprehensive systematic review of estimates of HbA1c variation includes data from 111 studies. I provide separate estimates of HbA1c variation in healthy and diabetic populations. I also observe a positive correlation between mean HbA1c and within-individual variation.

In the next chapter I report the systematic review describing another laboratory test, the inflammatory biomarker CRP.

3 Chapter three: Within-individual variation of C-reactive protein and high-sensitivity C-reactive protein: A systematic review and meta-analysis

A version of this chapter has been accepted for publication in PLOS One.

3.1 Abstract

Background

C-reactive protein (CRP) and high-sensitivity C-reactive protein (hsCRP) are measures of inflammation used in diagnosis, to guide treatment decisions, and in disease prediction. Variation in measured CRP and hsCRP may affect their clinical utility but estimates of within-individual variation are based on limited data.

Methods

A systematic review and meta-analysis was performed to estimate within-individual variation of CRP and hsCRP over any time period. Follow-up studies of any design in adults or children, with repeated measures of CRP or hsCRP were sought. Multiple databases were searched from inception to November 2022. Titles and abstracts were screened in duplicate. Full text screening and data extraction were performed by one reviewer and verified by a second. Risk of bias was assessed with a modified Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) tool. Intraclass correlation coefficient (ICC) results were pooled with a meta-analysis and coefficient of variation (CV) results were described by median and range.

Results

Of 2,675 studies identified, 60 met the inclusion criteria: 34 reported CRP and 26 reported hsCRP. For CRP, median CV was 0.41 (range 0.11 to 0.89), and the pooled estimate of ICC was 0.55 (95% CI 0.35 to 0.74). For hsCRP, median CV was 0.44 (range 0.27 to 0.76) and the pooled estimate of ICC was 0.62 (95% CI 0.58 to 0.67).

Limitations

Assessment of variation was not the main aim of many of the included studies, and it is possible that some relevant studies have been missed. Many of the studies included had low numbers of participants and/or low numbers of repeated measurements.

Conclusions

Estimated within-individual variation is high for both CRP and hsCRP, but estimates are based on small numbers of participants and measurements. There is a need for better estimates of within-individual variation from analysis of larger numbers of repeated measurements in larger numbers of subjects.

3.2 Introduction

C-reactive protein (CRP) is a non-specific biomarker for acute inflammatory diseases such as infections (including COVID-19) and for chronic inflammatory conditions such as cardiovascular diseases [155]. Plasma levels may increase up to a thousand-fold in the presence of inflammation and tissue trauma [156]. Mild, chronic elevations are associated with an increased risk of cardiovascular disease [157]. Point of care (POC) CRP can be used to aid the diagnosis and inform antibiotic treatment of lower respiratory tract infections (LRTIs) [158]. CRP is therefore used to inform diagnosis, prognosis and aid in therapeutic decisions making.

The CRP test evaluates a higher range of values than hsCRP. The CRP test is used to diagnose acute inflammatory conditions such as infection and to diagnose or monitor chronic inflammatory conditions like rheumatoid arthritis. The hsCRP test uses immunoassay methods which are sufficiently precise to quantify CRP throughout its normal range [157]. hsCRP helps predict risk of cardiovascular disease. Its most important use is to identify patients with known atherosclerotic cardiovascular disease with residual risk [159].

Various cut-off points are used when using CRP as an aid to diagnosis. A value of 0 to 0.3mg/L is considered normal, 1-10mg/L suggests moderate inflammation (autoimmune/ inflammatory diseases), 10-50mg/L suggests marked inflammation such as acute infection and more than 50mg/L suggests bacterial infection. A value less than 10mg/L suggests acute infection is unlikely [160]. Different researchers recommend different exact cut-offs. For example some suggest less than 10mg/L is normal and above 10mg/L indicates pathology [161]. NICE guidelines advise that

bacterial lower respiratory tract infection is unlikely with a CRP <20mg/L [162].

Although the reference range for hsCRP is considered to be 0 to 0.3 mg/L, this may vary in different settings. Higher average values have been observed in healthy adults in both Iran and South Korea [163, 164].

Reference ranges for diagnostic tests are usually derived from a reference population, which allows assessment of between-individual variation. However, biological variables also vary within individuals over time. This can be systematic and predictable, such as circadian or seasonal variation, or may be due to chance. This type of within-individual variation (CV_I) is known as biological variation. Variation can also be introduced into a measurement from pre-analytical factors. For example, transient infections, physical activity and stress increase CRP levels [165, 166] and CRP also increases with age [167]. Imperfect accuracy and precision of a test suggest that if a test is repeated a number of times, there will be a range of results around the true value. This is analytical variation (CV_A). Analytical and biological within-individual variation, CV_A and CV_I , together with any pre-analytical factors affecting the measurement, combine to give the total within-individual variation (CV_T). The coefficient of variation (CV), which is the standard deviation divided by the mean, is the most common statistic used to describe within-individual variation, but various other statistics are sometimes used, especially standard deviation (SD) and intraclass correlation (ICC).

Measured CRP results may differ from the true mean, i.e. the real mean value of CRP in an individual. This can be due to a combination of biological, pre-analytical and analytical variation. This total variation is the variation encountered in real-world clinical practice. Biological variation is often the greatest component of total variation, since clinical laboratories usually set acceptable analytical variation at <0.5% of

biological variation [15]. Different clinical processes and statuses have different levels of biological variation. For example, one study estimated the within-individual coefficient of variation (CV_i) of creatinine as 0.064 in people without chronic kidney disease [168] and another study estimated the CV_i of glucose in non-diabetics as approximately 0.05 [144].

CRP is considered to have a high biological variation. The European Biological Variation Database lists the CV_i of CRP as 0.341 (95%CI 0.294-0.744) and hsCRP as 0.589 (0.532-0.660) [77]. It has been argued however that there may be inaccuracies in published reports, for example because of inclusion of subjects with transient illness. In theory values higher than 0.33 for CV_i should be treated with suspicion [165].

The greater the within-individual variability of a biological process or status, the lower the probability that a single measurement is an accurate reflection of the individual's true mean. This can lead to errors in diagnosis, prognosis and treatment. Specifically for CRP, an inaccurate result may lead to inappropriately prescribing or withholding antibiotics in cases of sepsis and respiratory tract infections. An inaccurate hsCRP result may give inaccurate diagnoses and risk assessments for cardiac diseases. It has been shown that CRP values affect clinical decision-making. When paediatricians were given case vignettes with identical radiographs and clinical data, but randomised so that they were given CRP values either low (23mg/L) or high (151mg/L), they were 60-90% more likely to diagnose bacterial pneumonia rather than viral pneumonia if given the high CRP value, and consequently to prescribe antibiotics [169].

Some authors even suggest the level of variability of CRP may influence clinical decisions, with one paper studying intensive care unit acquired infections showing daily CRP variation could be used as a marker of infection [170].

A scoping review searching for studies on within individual variation of CRP (see Appendix 10,1, 20.1.2) found very little existing data on the subject, and a search of Pubmed, Epistemonikos and the Cochrane database revealed no previous large scale systematic reviews of within-individual variation of spirometry measures. The meta-analyses on the European Biological Variability Database [77] are based on only four studies for CRP and one paper for hsCRP. Note that this pragmatic review allows the inclusion of real-world data to improve the clinical applicability of the results.

3.3 Aim of review

To describe and evaluate the current literature on the within-individual variability of CRP and hsCRP in healthy subjects and patients with medical conditions and to estimate within-individual variability of these measures.

3.4 Methods

PRISMA and PRISMA-S guidelines were followed in the reporting of this review [134]. The review protocol was registered at Prospero (CRD42020202387 Registered 17/09/2020). The review protocol is available at https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020202387

The original study intended to include HbA1c and CRP in a single paper, but for reasons of size and to improve targeting of appropriate outlets for dissemination, the study was separated into HbA1c and CRP after the data extraction stage.

The methods for this systematic review were as described in Chapter 2 for the HbA1c systematic review, except as noted below. The searches for this review were combined with the systematic review of HbA1c variability, (with dates of searches and databases searched given in Chapter 2.4) and then HbA1c studies were excluded at the full text screening stage.

Studies were included if primary research data on the variability of at least two measurements of CRP or hsCRP within the same individual were recorded in order to include as many studies as possible.

Table 3.1 lists the outcome and other main variables extracted. Variables selected for extraction and for subgroup analysis were those I hypothesised as most likely to affect variability such as health status and age. The primary outcome measure was variability of repeated measurements within the same individual. If both CRP and hsCRP were included, the data were extracted separately for each measurand.

Table 3.1 Variables extracted from studies

Variable	Definition and rules
CRP or hsCRP	Whether high sensitivity CRP was specified in the text
Study design	Cohort, RCT etc
Number of subjects	Number of subjects used to calculate variability measure
Age	Average age of subjects used to calculate variability measure. If the subjects used to calculate variability measure is a subset of the study population and only the age is given for the whole population, the age of the whole population is used.
Sex	Percentage of subjects used to calculate variability measure who were male. If the subjects used to calculate variability measure is a subset of the study population and only the sex is given for the whole population, the sex of the whole population is used.
Ethnicity	Ethnicity recorded yes/no? If country of origin of subjects only is recorded, this is counted as no.
Setting	Primary care/community versus secondary/tertiary/laboratory setting
Health status	Healthy v inflammatory disease v cardiac disease v haemodialysis v other
Number of measurements	Number of repeated CRP measurements in the same individual
Time interval between measurements	Length of time between measurements in days.
CV	Coefficient of variation of repeated measures within an individual
SD	Standard deviation of repeated measures within an individual
ICC	Intraclass correlation of repeated measures within an individual

A risk of bias tool was adapted from the COSMIN tool for test reliability [136]. Risk of bias was also assessed using the BIVAC method, a tool developed by the EFLM to assess the quality of biological variation studies [154]. See Appendix 3.3.

Studies were grouped for synthesis based on test (CRP vs hsCRP) and variability measure reported e.g. ICC, CV. hsCRP and CRP measurements were analysed separately. Units of CRP measurement were converted to mg/L.

3.5 Results

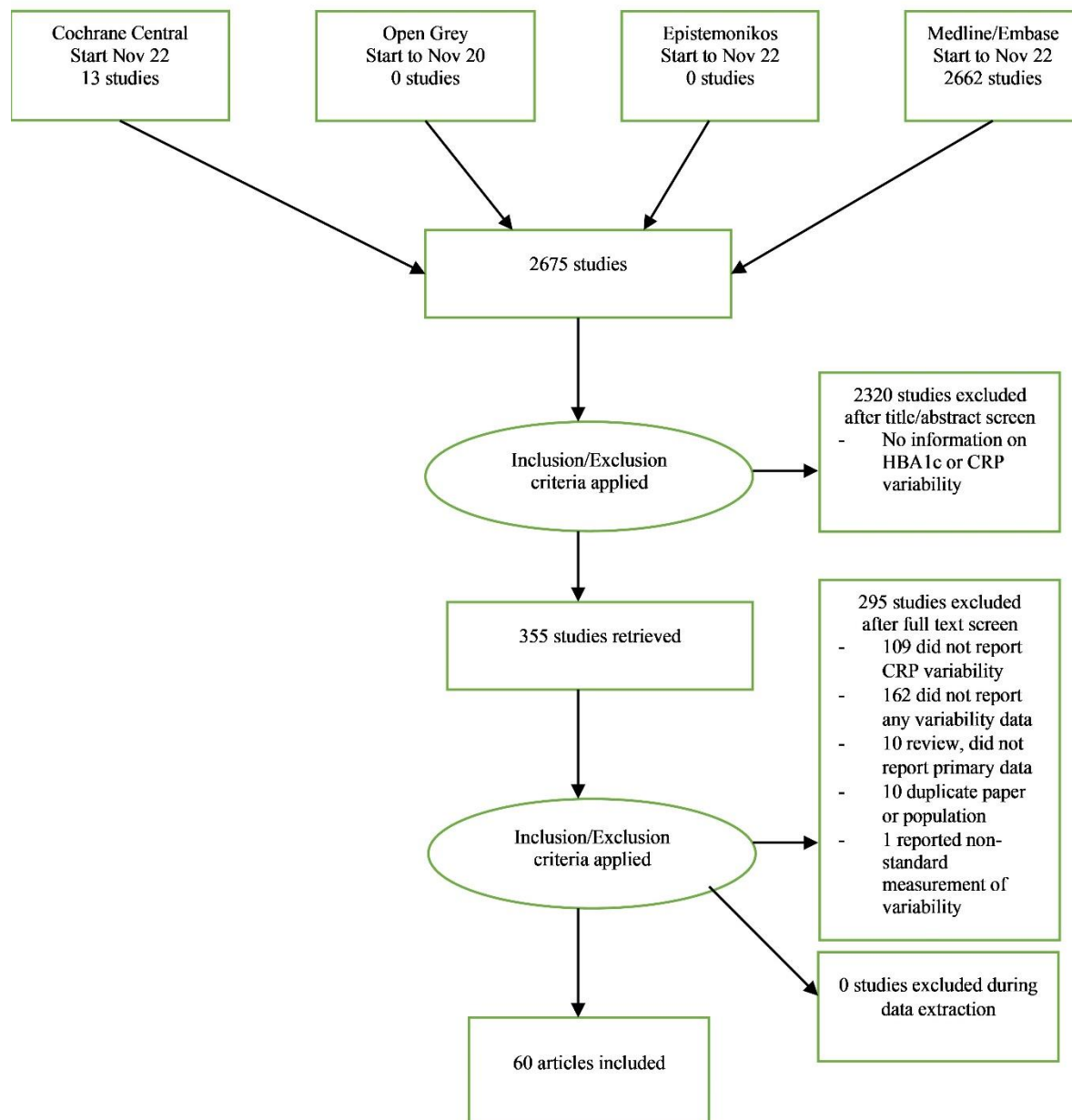


Figure 3.1 PRISMA flow diagram describing selection of studies

After database searches and hand searching of reference lists of included studies, 2,675 non-duplicate studies were retrieved, and 2,320 studies (86.7%) were excluded after title and abstract screening leaving 355 (13.3%) articles. Studies were

excluded at the title and abstract screening stage if it was clear there was no information on variability of either HbA1c or CRP in the study. After full text screening, 295 (83.1% of full texts screened) studies were excluded. Of these, 109 (36.9%) were excluded from this study because they reported variability for HbA1c only and not CRP, 162 (54.9%) were excluded because they did not report any variability data, 10 (3.4%) were excluded because they did not report any primary data (e.g. they were a review study), 10 (3.4%) were excluded because they were a duplicate paper or population, and 1 (0.3%) was excluded because it reported a non-standard measurement of variability (the mean absolute residual around the line connecting index value with closing value) (see fig 3.1).

Sixty (16.9% of full texts screened) studies met the inclusion criteria, including between four and 56,218 subjects. Fifty-four studies (90% of included studies) were of a cohort design and six (10%) used data from the placebo arm of a randomised controlled trial. Fourteen studies (23.3%) were carried out in a primary care or community setting and 38 (63.3%) studies were carried out in secondary or tertiary setting such as universities, hospitals and laboratories. In eight (13.3%) studies the setting was not reported.

Thirty-four (56.7%) studies reported CRP results, and 26 (43.3%) reported hsCRP results. Fifty-one (84%) studies reported long-term variability (arbitrarily defined as 7 or more days between measurements), six (10%) reported short-term variability (less than 7 days between measurements), two studies reported both long- and short-term, and one did not report measurement intervals. Study populations were diverse in terms of age and gender (see table 3.2). Forty studies (66.7%) were in healthy populations, 8 (13.3%) were in populations with respiratory disease, 3

(5%) were in populations with heart disease, 6 (10%) were in populations undergoing haemodialysis and 3 (5%) were in populations with other health statuses.

Ethnicity was recorded in 11 (18.3%) studies. In 22 (36.7%) studies the within-individual biological variation (CV_I) was reported, while in 35 (58.3%) studies the within-individual total variation (CV_T) was reported. In three (5.0%) studies it was uncertain whether total or biological variation was reported. For CRP, the number of studies reporting CV_I , CV_T or it was uncertain were 14/34 (41.2%), 18/34 (52.9%) and 2/34 (5.9%) respectively. For hsCRP the number of studies reporting CV_I , CV_T or it was uncertain were 8/26 (30.7%), 17/26 (65.4%) and 1/26 (3.8%) respectively.

Aims of the studies included explicitly measuring biological variation, investigating associations between biological variation and diseases or outcomes such as cognitive decline, polycystic ovary syndrome and atherosclerosis, identifying risk factors associated with increased biological variability such as COPD, and calculating reference change values.

Table 3.2 shows summaries of the study characteristics and results. See Appendix 3.2 for full characteristics of included studies.

Table 3.2 Summary statistics of study characteristics and results of included studies

Study characteristics	Number of observations	Median	IQR	Min	Max
Number of subjects	59	50	139	4	56218
Age (years)	40.0	49.5	33.1	18.0	72.9
% male	51	50	42	0	100
Number of measurements	58	4	6	2	18
Study results					
CV	32	0.421	0.183	0.106	0.894
SD	7	0.80	1.93	0.04	2.04
ICC	19	0.62	0.20	0.17	0.90

For CRP there were 21 studies (35.0% of all studies) that reported a CV, with a range of 0.11 to 0.89 and a median of 0.41. For hsCRP there were 11 studies (18.3%) that reported a CV with a range of 0.27 to 0.76 and a median of 0.44. For CRP there were seven studies that reported an ICC. The pooled estimate calculated in this study for CRP ICC was 0.55 (95% CI 0.35 to 0.74). For hsCRP there were 12 studies that reported an ICC. The pooled estimate calculated in this study for hsCRP ICC was 0.62 (95% CI 0.58 to 0.67).

Figures 3.2 and 3.3 show forest plots of all studies that reported a CV split by health status and sorted by number of subjects for CRP and hsCRP respectively. Results were heterogeneous for the healthy and unhealthy groups, making it hard to draw conclusions.

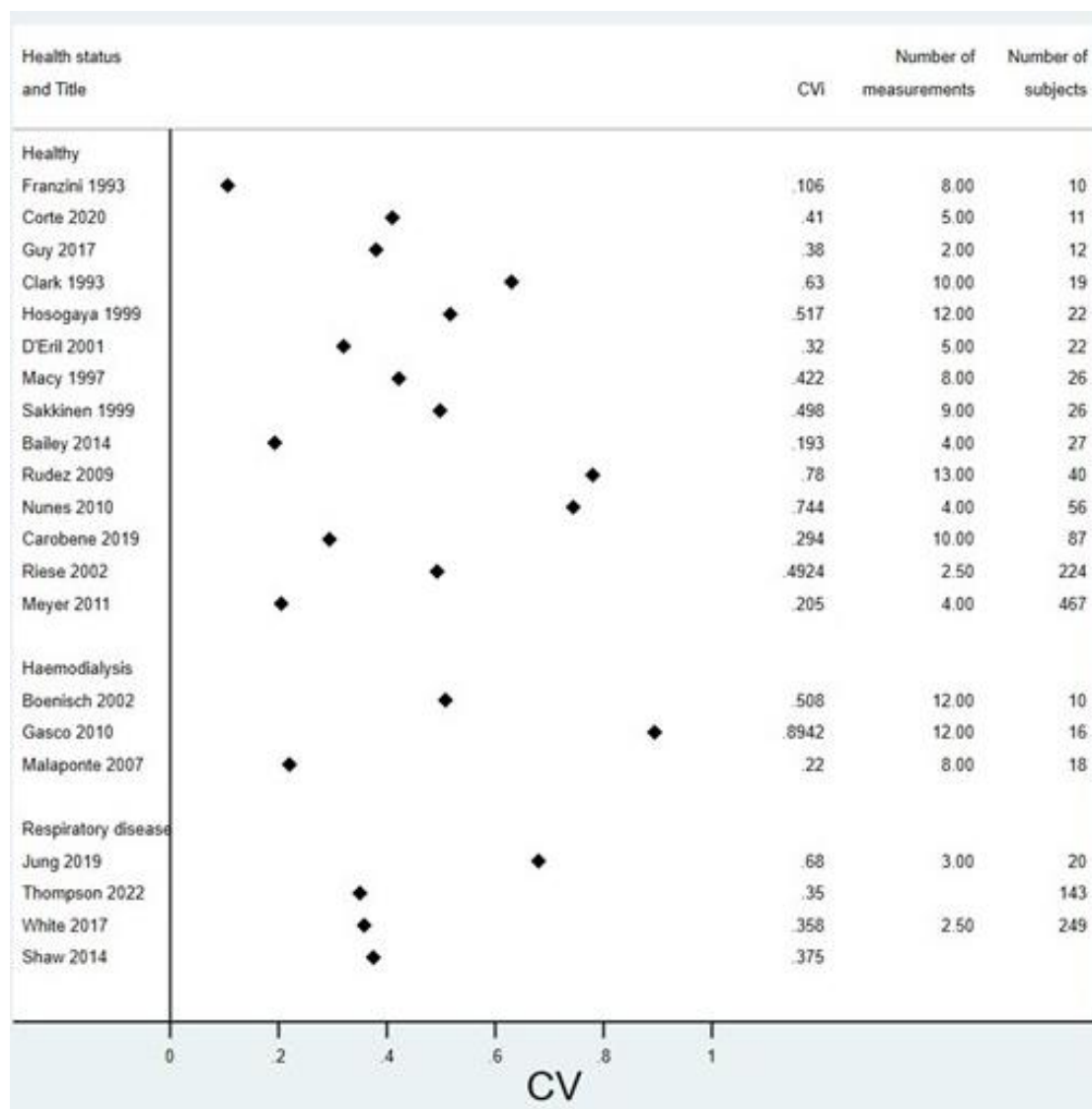


Figure 3.2 Modified forest plot of all studies that reported a CV for CRP, split by health status

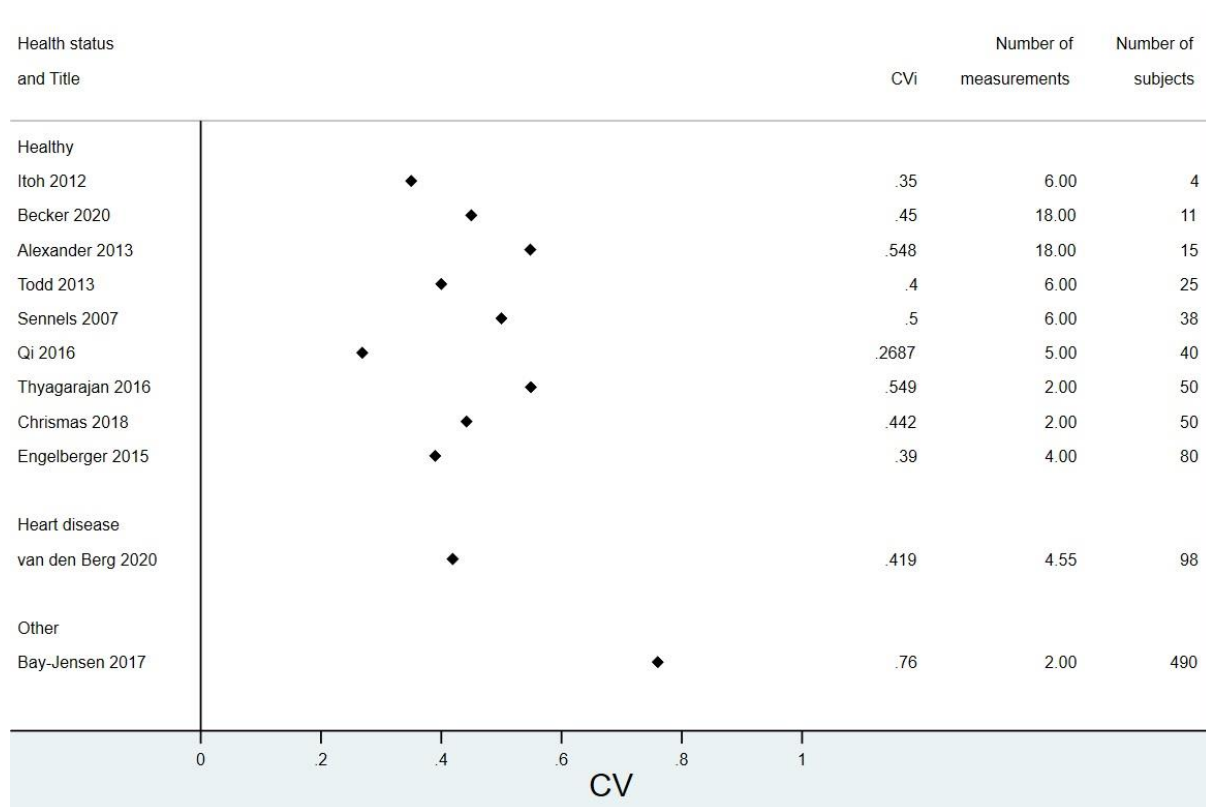


Figure 3.3 Modified forest plot of all studies that reported a CV for hsCRP, split by health status

Figures 3.4 and 3.5 show forest plots with meta-analysis of all studies that reported an ICC, split by health status and sorted by number of subjects for CRP and hsCRP respectively. Sizes of subgroups were too small to draw conclusions about differences between the groups.

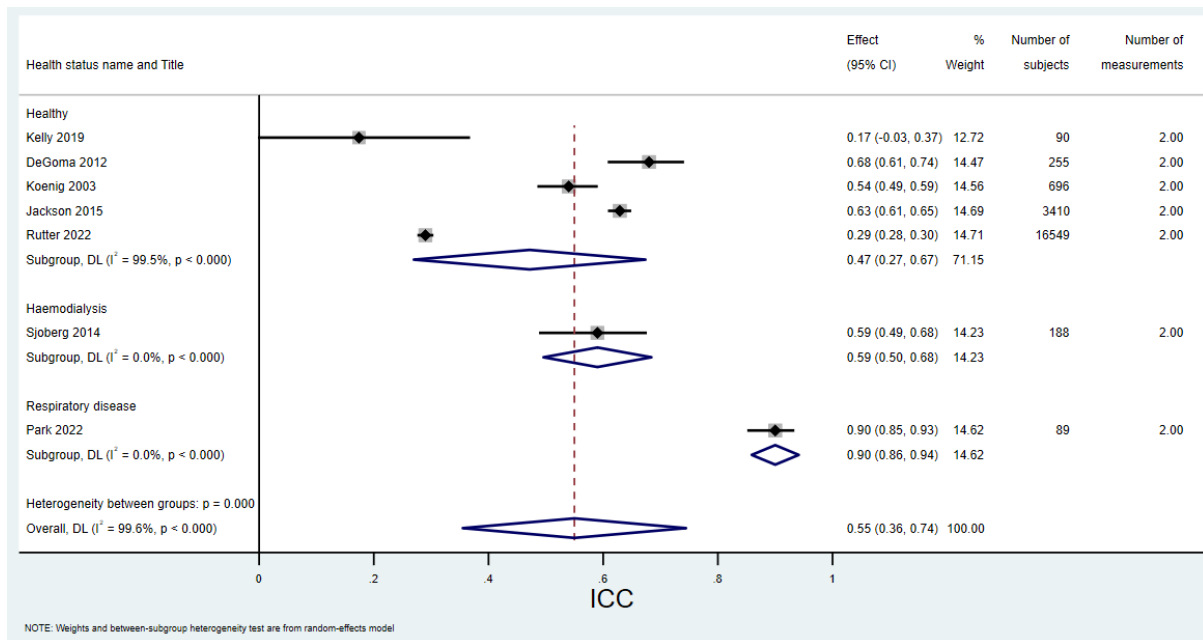


Figure 3.4 Forest plot of all studies that reported an ICC for CRP, split by health status, with meta-analysis.

Note that the lower confidence interval for Kelly was manually truncated at 0, since the meta-analysis gave a negative value which is meaningless for an ICC.

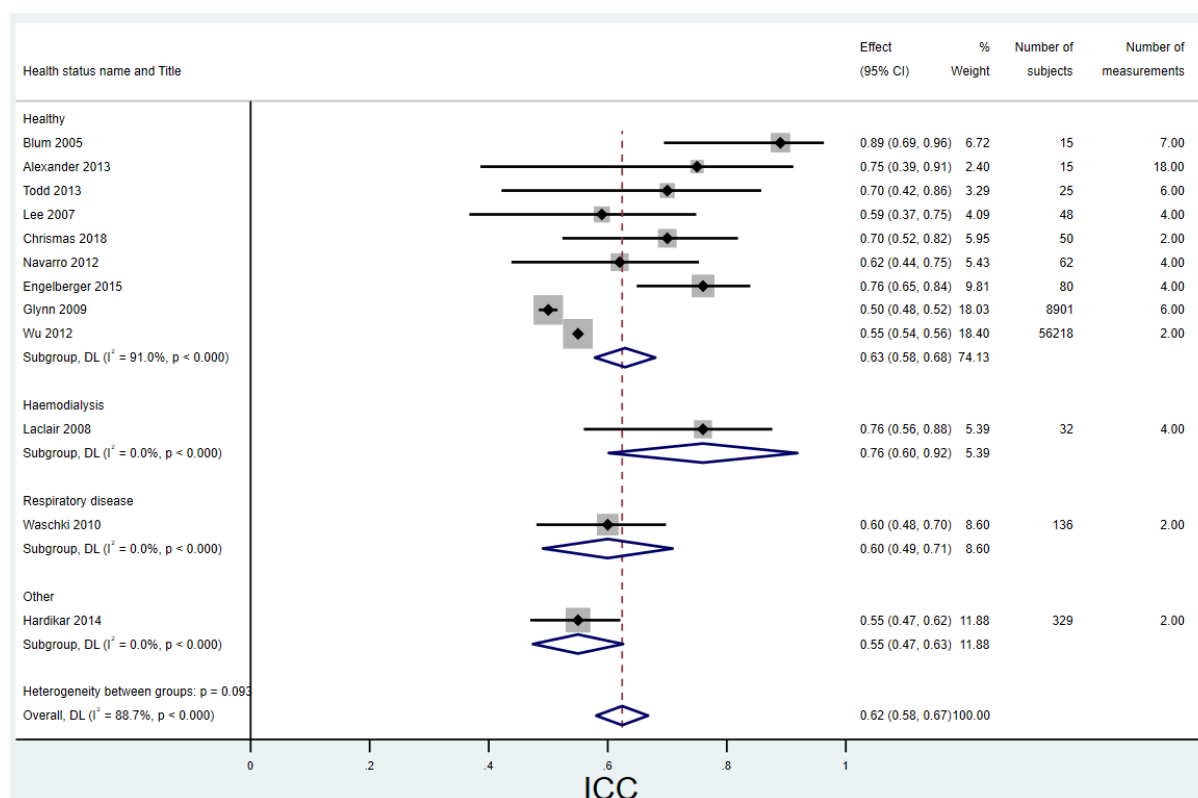


Figure 3.5 Forest plot of all studies that reported an ICC for hsCRP, split by health status, with meta-analysis

Figures 13.1 to 13.4 show forest plots of CVs for CRP categorised by risk of bias, setting, short- or long-term variability and unit of measurement respectively. Figures 13.5 to 13.7 show forest plots of CVs for hsCRP categories by risk of bias, setting and short- or long-term variability respectively. Figures 13.8 and 13.9 show forest plots with meta-analysis categorised by risk of bias for ICCs of CRP and hsCRP respectively. There was little difference between the subgroups. See Appendix 3.4 for figs 13.1 to 13.9.

When assessing risk of bias, no study scored more than doubtful for question B4 *knowledge of previous scores when administering the test* and B5, *when assigning the scores* (questions B4 and B5). Table 3.3 shows the frequency of the maximum risk of bias with questions B4 and B5 excluded since including them would

mean that all studies would be rated at a high risk of bias. Appendix 3.3 shows the risk of bias of all included studies. The inclusion criteria ruled out inadequate studies.

Table 3.3 Highest risk of bias with questions B4 and B5 excluded.

Maximum risk of bias	Number of studies (%)
1	38 (63.3)
2	15 (25.0)
3	7 (11.7)
4	0 (0)

(1 is best possible score, 4 is worst possible score)

Risk of bias due to missing results was deemed to be low, since in most cases the outcome measure of this review was not the primary outcome of the individual studies, and therefore withholding publication based on unwanted results is unlikely. Some studies with repeated measures do not report variability data, but this is likely to be because it is not of relevance to the study. As such this should happen at random, rather than introducing a systematic bias. If studies with missing results had missing data systematically related to the variability measures, this could introduce bias, but there is no reason to think this is the case.

3.6 Discussion

This review describes the current literature on variation of CRP and hsCRP within an individual. While it is recognised that CRP and hsCRP have high biological variation, the data regarding this is spread throughout the literature, with the only meta-analysis on the subject based on very low numbers of studies. This current systematic review and meta-analysis has a more transparent and consistent search strategy than the existing database. It includes a much wider pool of studies, and is not restricted to healthy populations, allowing some conclusions to be drawn on the

variation of CRP and hsCRP in unhealthy populations. Studies reported a wide range of variation.

For CRP, the median CV was 0.41 with a range of 0.11 to 0.89 and the pooled estimate ICC was 0.55 (95% CI 0.35 to 0.74). For hsCRP the median CV was 0.44 with a range of 0.27 to 0.76 and the pooled estimate ICC was 0.62 (95%CI 0.58 to 0.67). These compare with the pooled results for CV_i given on the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Biological Variation Database [77] of 0.34 (95%CI 0.29 to 0.74) for CRP and 0.59 (95%CI 0.53 to 0.66) for hsCRP. This likely reflects a wide range of health statuses, settings, time between measurements and number of measurements. The EFLM database largely includes studies on healthy subjects. However, subgroup analyses and sensitivity analyses in this study showed little difference between subgroups.

The paper by Kelly [171] is an outlier with no obvious explanation for the very low ICC. The paper also examined within-individual variation of HbA1c and reported a much higher ICC for HbA1c of 0.873. The study was unremarkable in terms of setting and patient characteristics. However, the paper was a conference abstract with no full text, so further detail and discussion of this result was lacking.

There is no standard agreement for what level of coefficient of variation or intraclass correlation constitutes “low” or “high” variation, and in fact this needs to be individualised for each measurand. A measurand that has a large within-individual coefficient of variation, but a large disparity in measured values between healthy and diseased populations may have less clinical implication than a lower level of variation but a lower disparity between healthy and diseased values. Nevertheless, the level of variation for CRP and hsCRP shown in this systematic review is subjectively high.

In theory the CV of a positive normally distributed set of data should not exceed 0.33 (11) which suggests that either the range of values of CRP is not normally distributed or that some of the values are incorrect due for example to transitory illnesses. However, there seems to be sufficient data in the current literature to suggest that whatever the “true” figure for CRP variation, it is high enough to be clinically significant. This is particularly true when single number cut-off points are used, for example “do not use antibiotics if CRP <20mg/L”, since a measured value of for example 18mg/L may represent a much higher (or much lower) true mean for that individual. While CRP alone is not the basis for clinical decisions on treatment with antibiotics, since many factors besides infection can affect CRP levels, if using CRP measurement results to aid diagnostic decisions, clinicians should be aware that a single measurement, or even multiple measurements, may not be an accurate reflection of the true mean CRP for an individual patient. Within-individual variation may also be important with regard to hsCRP levels, particularly in the setting of cardiovascular disease.

The studies included in this review often had low numbers of participants and low numbers of repeated measurements. The studies with high numbers of participants tended to have low numbers of repeated measurements and vice versa, likely for practical and cost reasons. There is thus a lack of studies with a large number of repeated measurements in a large population. None of the studies were blinded, (it was not ensured that the professionals administered the tests and assigned the scores without knowledge of previous patient scores).

The effect of setting on variation of laboratory tests is discussed in Chapter 2.6.

There are some limitations with this review. Not all studies that report variation have the primary aim of measuring variation, so it is possible that some studies with data on CRP variation may have been missed. However, a broad range of search terms was employed which retrieved a large number of titles and abstracts, which in turn led to a large number of full texts being screened for inclusion.

For meta-analysis, only a small number of studies could be pooled because of different tests being performed (CRP v hsCRP), different variation measurements being calculated (CV, SD and ICC) and lack of well-described validated methods for performing meta-analysis of coefficients of variation. It should also be noted that some studies reported total within-individual variation (i.e analytical plus biological variation) and some biological within-individual variation, and in some cases an assumption was made as to which of these had been calculated. In a few cases it was not even possible to make an assumption as to which type of variation was being described (see Appendix S5).

The inclusion of data from the placebo arm of randomised controlled trials could be considered to be problematic if placebo has an effect on the variation. However, I am not aware of any evidence that this effect can occur.

Limitations with the risk of bias tools are discussed in Chapter 2.6.

Note that some studies may not have clearly distinguished between a material change in the disease in populations that have the disease (e.g. an increase in mean value) and an increased level of variation due to the disease (the two aspects of pathological variation), which could increase the measured variability and influence the association of increased variability with presence of disease/increased disease severity.

Note that some of the underlying conditions for which CRP is part of the diagnostic or monitoring pathway are more variable than other conditions (e.g. bacterial pneumonia versus rheumatoid arthritis), and also more variable than other underlying conditions for which the tests examined in this thesis are performed.

Further studies are required that have both larger populations and more repeat measurements, in a variety of different health statuses, to enable the variation of CRP and hsCRP to be better described. Future research should consider standardisation of reporting of measures of variation such as using the coefficient of variation, and making sure that the values for within-individual variation are clearly reported. Standardised biological variation studies such as those listed in the ELFM database generally use healthy populations, but it is important that research is directed at ascertaining the within-individual measured variation in populations with various disease statuses. The implications of within-individual measured variation on reference ranges, the formulation of clinical guidelines and clinical decision-making need to be investigated.

Nevertheless, this systematic review and meta-analysis provides new information on the within-individual variation of CRP.

3.7 Conclusion

The high variation shown by this review demonstrates the need for clinicians and authors of clinical guidelines to carefully consider how accurately a single CRP result reflects the individual's true mean CRP level, and how many repeat tests may be needed to provide an accurate result.

In the next chapter I report the systematic review of a non-laboratory test, since most of the work on within-individual variation that has been done to date has involved laboratory tests, mainly biochemistry tests. It was therefore decided to look at the extent and quality of the data on within-individual measured variation in the spirometry tests FEV1 and FVC which are commonly measured in primary care settings. This is a type of measurement which differs from laboratory measurements because there is some degree of observer error.

4 Chapter four: Within-individual variation in respiratory function tests - A Systematic Review and Meta-analysis

4.1 Abstract

Objectives

Within-individual variation in respiratory function tests affects their clinical utility. Chance within-individual variation causes measured results to differ from the patient true mean, and this may have clinical implications. No systematic review has described the within-individual variation of FEV1 and FVC.

Methods

Research databases (Medline, Embase, Cochrane, Epistemonikos, Open Grey, Google Scholar) were searched using a predefined search strategy from inception to 9th July 2021 for studies reporting within-individual variation of Forced Expiratory Volume in one second (FEV1) or Forced Vital Capacity (FVC) in any participants. Titles and abstracts were screened independently by two reviewers. Full texts were screened by one reviewer and exclusions confirmed by a second reviewer. The data extraction for 50% of the studies were checked by a second reviewer. Risk of bias was assessed using a modified consensus-based standards for the selection of health measurement Instruments (COSMIN) tool. Descriptive

statistics for coefficients of variation (CVs) were reported and a meta-analysis of Intraclass Correlation Coefficients (ICCs) was performed.

Results

104 studies met the inclusion criteria. Of these, 45 (43.3%) were on healthy populations, 59 (56.7%) in patients with: asthma 18 (17.3%), COPD 15 (14.4%), cystic fibrosis 8 (7.7%), multiple health statuses or other conditions 18 (17.3%).

The median CV (within-individual coefficient of variation) of FEV1 was 0.034 (IQR 0.025 to 0.067) and of FVC was 0.031 (0.020 to 0.048). In healthy populations median CV of FEV1 was 0.028 (range 0.011 to 0.177) and of FVC was 0.029 (range 0.010 to 0.163). In COPD: FEV1 was 0.068 (range 0.020 to 0.122) and FVC 0.068 (range 0.016 to 0.140). In asthma, FEV1 was 0.030 (range 0.018 to 0.089) and FVC 0.024 (range 0.014 to 0.029).

Discussion

Many included studies had few participants or few repeated measurements. Within-individual variation of FEV1 and FVC is higher for patients with than without COPD.

4.2 Aim of review

To describe and evaluate the current literature on the within-individual variation of FEV1 and FVC in healthy subjects and patients with medical conditions and to estimate within-individual variation of these measures.

4.3 Introduction

Spirometry is the measurement of air volumes during expiration which allows an estimation of lung function and capacity [172]. It is important in the diagnosis, classification and management of respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) [173]. Spirometry is recommended for the diagnosis of COPD in adults and asthma in children [174] [175]. Two of the most commonly evaluated measurements obtained during spirometry are the Forced Expiratory Volume in one second (FEV1), that is the volume of air expelled under force in the first second, and the Forced Vital Capacity (FVC), that is the maximum volume of air that can be forcefully expelled [176]. These can be reported in absolute values (litres or mLs), or in relative values (percentage of predicted value) [177, 178]. Reference ranges vary by height, sex, age and ethnicity [177, 178]. A FEV1/FVC ratio of <0.7 is consistent with a diagnosis of airflow obstruction. Reversibility testing with bronchodilators can help distinguish between COPD and asthma but are not completely reliable [179]. Predicted FEV1 value can be used to categorise disease severity, with NICE guidelines classifying $>80\%$ as mild, 50-80% as moderate, 30-50% as severe and $<30\%$ classified as very severe [180].

Reference ranges set by comparison to a reference population allow assessment

of between-individual variation. However, biological processes such as lung function also vary over time. This can be systematic and predictable, such as seasonal or circadian variation. For example, spirometry results have been shown to vary by time of day [181]. However, variation may also be due to chance. Within-individual variation shows variation around a mean, which is distinct from decline over time, where the mean gradually reduces. Within-individual variation can be divided into pre-analytical, analytical and biological variation. Pre-analytical variation refers to changes in the individual or the measurement method prior to the measurement being carried out. For example, pre-analytical variation can be caused by recent exercise, diet changes or stress. Analytical variation refers to the inherent imprecision within the testing method. In the case of spirometry, factors causing analytical variation may include calibration errors, system leaks and environmental factors such as humidity and temperature [182]. Biological variation refers to the changes in the measurement over time due to intrinsic factors which may be predictable such as pathological and physiological changes or may be due to random chance.

In the case of spirometry, within-day and especially within-session variation may approximate to the sum of pre-analytical and analytical variation. In laboratory medicine, a sample can be run in duplicate and the difference compared in order to assess analytical variation but this is not as simple in spirometry, since each result is generated by an individual, with different levels of airflow being produced due to variation in individual motivation, fatigue or technical correctness of carrying out the test. Pulmonary waveform generators which simulate the spirometry patterns of healthy individuals and patients with respiratory disease can be used to measure instrument accuracy/analytical variation [183]. For a syringe with $\pm 0.5\%$ accuracy, a

$\pm 3.5\%$ volume accuracy is considered appropriate [184]. However, this review included only results from humans. For a clinical assessment, three acceptable spirograms must be obtained, and the largest two for both FEV1 and FVC must be within 0.15L of each other [184]. This would have the effect of reducing within-session/within-day variation. Between-day variation approximates to the sum of the pre-analytical, analytical and biological variation.

It has been reported that lung function measurements tend to be more variable when made weeks or months apart compared to measurements made on the same day [185]. Spirometry measurements taken temporally close to each other, e.g. within a day, when the underlying disease process is unlikely to have fluctuated significantly, may approximate to the pre-analytical variation and analytical variation. Measurements taken further apart, but sufficiently close in time not to be impacted by disease deterioration (i.e. lung function decline) evaluate the total variation. The greater the within-individual variation of a measurand, the lower the probability that a single measurement (or in the case of spirometry the best of three blows) is an accurate reflection of the true mean of the measurand in that individual. This can lead to errors in diagnosis, prognosis and treatment. It is suggested that serial spirometric measurements can be useful to help with the assessment of progression and management of the disease (3). However, if chance within-individual variation is not taken into account, this could lead to inappropriate clinical decisions. For example, it has been shown in blood pressure testing that not accounting for within-individual variation can cause misclassification of severity [43]. It is therefore important to understand and quantify within-person variation because it affects the interpretation of a change in FEV1 and FVC. Knowledge of the level of variation can

help to distinguish a true change in patient mean FEV1 and FVC from chance variation in measured FEV1 and FVC.

A scoping review searching for studies on within individual variation of spirometry measures (see Appendix 10,1, 20.1.7) found very little existing data on the subject, and a search of Pubmed, Epistemonikos and the Cochrane database revealed no previous systematic reviews of within-individual variation of spirometry measures. This study therefore aims to describe and evaluate the current literature on the within-individual variation of FVC and FEV1 in healthy individuals and patients with physical disorders and to estimate within-individual variation of these measures.

4.4 Methods

PRISMA and PRISMA-S guidelines were followed in the reporting of this review [134]. The methods for this systematic review were as described in Chapter 2 for the HbA1c systematic review, except as noted below.

Searches were devised to identify studies in which a spirometry measurement was repeated in the same individual. See Appendix 4.1 for sample Medline and Embase search strategy.

The international prospective register of systematic reviews, PROSPERO, was checked for ongoing reviews, and the protocol was submitted to PROSPERO on 7th May 2020 (Prospero CRD42020184190). The review protocol is available at https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=184190

On 19th May 2020, database searches were carried out. Medline, Embase, Cochrane Central, Epistemonikos and Open Grey were searched from inception to 19th May 2020. Searches were updated on 9th July 2021.

Inclusion criteria for population included any age range, sex or setting, of any health status, with or without lung disease, and treated or untreated. Studies were included if a measure of the within-individual variation of FEV₁ or FVC was reported, with at least two measurements in the same individual performed. Primary studies of any trial design were included e.g. cohort studies, randomised controlled trials or case-control studies. There was no restriction on date of publication or language. Studies were excluded if participants were not in a stable state (measurements were before and after an intervention including bronchodilation or had an acute or rapidly changing illness) or data were secondary (i.e. reported from another study) to avoid duplication of populations.

In addition to studies identified with the search strategy, the following large cohort studies which were not identified by the searches but were suggested by subject experts (Turner, A. personal communication 2023) were checked: Hoesein [186], Casanova et al 2012 [187], Kohansal et al 2009 [188] and Woodruff et al 2016 [189]. None of these had data on within-individual variation.

Foreign language studies were translated by Google translate software. Data were extracted by a single reviewer, and 50% of the data extraction was verified independently by a second reviewer and the other 50% rechecked by the first reviewer, with disagreements between resolved by discussion. The decision not to have 100% of the data verified by a second reviewer was a pragmatic decision based on available resources.

Appendix 4.2 lists the outcome and other main variables extracted. Variables selected for extraction and for subgroup analysis were those I hypothesised to most

likely affect variability such as disease status and age. If both FEV1 and FVC data were reported, the data was extracted separately for each spirometry measure.

The main outcome measure of this review is within-individual coefficient of variation of FEV1 and FVC as the most common measure of variation reported in the types of studies included in this review.

Descriptive subgroup analyses were performed based on unit of measurement (Litres, millilitres or % predicted), health status, setting, number of measurements and whether short or long-term variation was measured.

Where FEV1 and FVC were reported in same study and patient characteristics differ between those groups, the patient characteristics of the FEV1 cohort are reported here. Studies were pooled to provide combined results where possible (i.e. the same spirometry measurement, FEV1 or FVC, and the same measure of variation).

4.5 Results

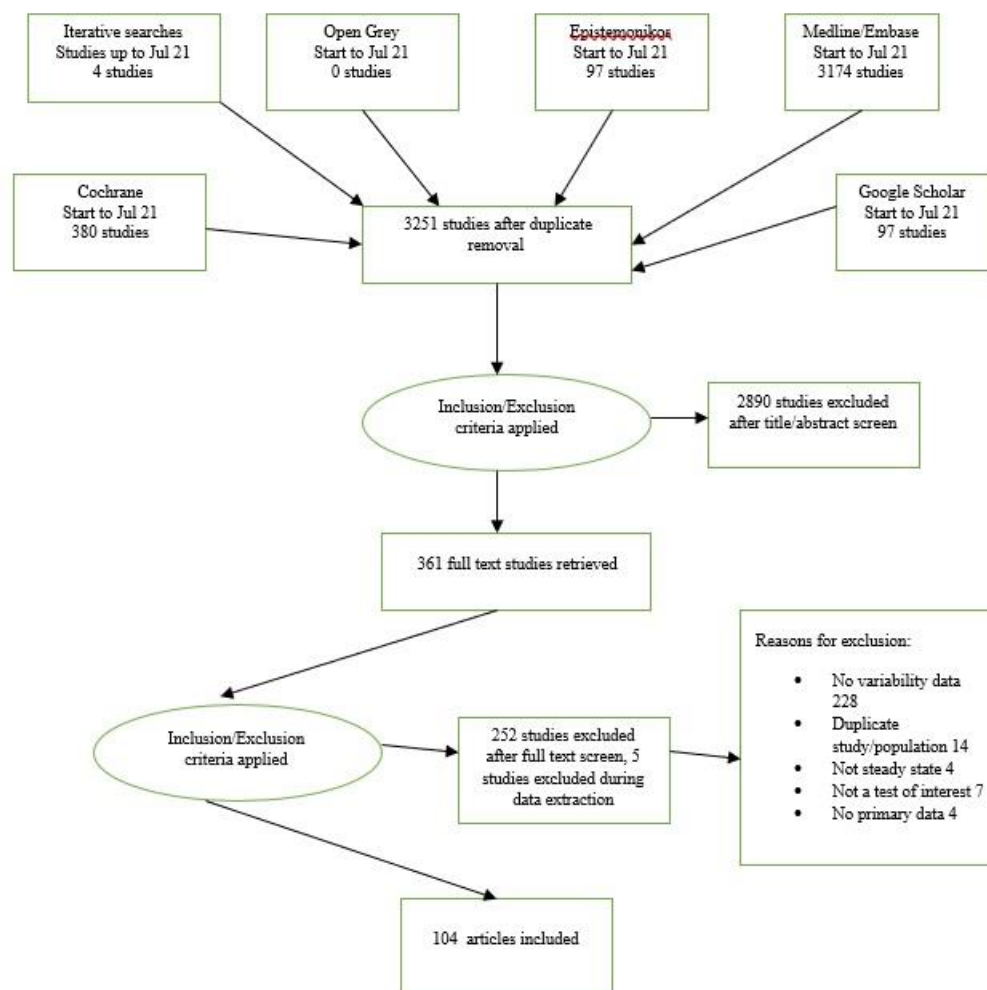


Figure 4.1 PRISMA flow diagram describing selection of studies.

Three thousand two hundred and fifty-one non-duplicate citations were retrieved after database searches of Embase, Medline, Cochrane, Epistemonokis and hand searching of reference lists of included studies. Two thousand eight hundred and ninety articles were excluded after title and abstract screening leaving 361 studies included. After full text screening, 252 studies were excluded, and a

further 5 were excluded during data extraction, leaving 104 included studies for analysis. 97 studies were cohort design, 6 were randomised controlled trials and one was a case control study. 85 of the studies were reported as full texts, and 19 as abstracts only (see fig 4.1). Most full text exclusions were due to lack of reporting of a recognised measure of variation.

Study populations were diverse in terms of age, sex and health status. See tables 4.1 and 4.2 for details.

Table 4.1 Characteristics of studies included (Paediatric populations).

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex % male (primary analysis)	Setting	Health status	Number of measurements	Days between measurements	Variation measurement interval	Risk of Bias score
Ranganathan 2002 [190]	Cohort	FVC	98	0.1		Other	Healthy	3	N/A	Within-day	1
Shin 2013 [191]	Cohort	Both	19	4.5	57.9	Primary	Healthy	2	90 to 90	Between-day	1
Vilozni 2001 [192]	Cohort	Both	32	4.5	43.8	Primary	Healthy	2	N/A	Within-day	3
Franca 2013 [193]	Cohort	Both	47	5.0	55.3	Primary	Healthy	2	21 to 28	Between-day	3
Strachan 1989 [194]	Cohort	Both	171	7.0		Primary	Healthy	2	7 to 28	Between-day	1
Fasola 2019 [195]	Cohort	Both	15	8.5	66.7	Other	Healthy	3	7 to 7	Between-day	1
Minsky 2016 [196]	Cohort	Both	58	9.2	39.7	Primary	Healthy	2	N/A	Within-day	3
Cooper 1990 [197]	Cohort	FEV1	23	13.2		Primary	Healthy	9	9 to 9	Between-day	1
Viegi 1988 [198]	Cohort	Both	15	17.5	100.0	Primary	Healthy	3	N/A	Within-day	3
Scalco 2018 [199]	Cohort	Both	22	9.3	36.3	Primary	COPD	3	1 to 14	Between-day	1
Mallol 2005 [200]	Cohort	FVC	102	1.0	71.6	Other	Asthma	4	N/A	Within-day	1
Wu 2009 [201]	Placebo arm of RCT	FEV1	417	8.5	59.7	Other	Asthma	10	120 to 320	Between-day	1
Timonen 1997 [202]	Cohort	Both	19	9.9	42.1	Primary	Asthma	12	7 to 7	Between-day	1
Kondo 1996 [203]	Cohort	FEV1	20	10.5	60.0	Other	Asthma	80	90 to 90	Between-day	1

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex % male (primary analysis)	Setting	Health status	Number of measurements	Days between measurements	Variation measurement interval	Risk of Bias score
Jean-Baptiste 2010 [204]	Cohort	Both	114	10.7	54.4	Other	Asthma	6	N/A	Within-day	3
Bastian-Lee 2002 [205]	Cohort	FEV1	27	11.0	41.0	Other	Asthma	2	N/A	Within-day	2
Edmondson 2020 [206]	Cohort	FEV1	74	7.0		Primary	CF	20	3-4	Between-day	3
Abbas 2013 [207]	Cohort	FEV1	25	11.3	40.0	Primary	CF	2	N/A	Within-day	2
Sanders 2008 [208]	Cohort	Both	40	15.3	47.0	Other	CF	3	N/A	Within-day	3
Iannaccone 2003 [209]	Cohort	Both	20	9.0		Other	Other	3	2 to 28	Between-day	1
Singer 2018 [210]	Cohort	FEV1	89	14.1	41.6	Other	Other	3		Between-day	3
Connolly 2015 [211]	Cohort	FEV1	91	16.7	100.0	Other	Other	3	N/A	Within-day	3
Florence 2012 [212]	Cohort	FVC	93	16.8	100.0	Other	Other	2	N/A	Within-day	3
Hagmolen Of Ten Have 2005 [213]	Cohort	FEV1	22	10.8	45.4	Other	Multiple	3	1 to 7	Between-day	1
Median			40	9.9	54.4			3			
Total			1653								

Setting :Primary=primary or community; Other=tertiary/secondary/laboratory. Risk of bias are presented as highest score, with questions 4 and 5 excluded. 1=very good; 2=adequate; 3=doubtful

Table 4.2 Characteristics of studies included (Adult populations).

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex %male (primary analysis)	Setting	Health status	Number of measurements	Range of time between measurements (days)	Variation measurement interval	Risk of Bias score
Nickerson 1980 [214]	Cohort	Both	15	18.0	46.7	Other	Healthy	4	N/A	Within-day	1
Kennedy 2017 [215]	Cohort	FEV1	11	20.2	45.4	Other	Healthy	2	1 to 1	Between-day	1
Ross 1992 [216]	Cohort	FEV1	14	23.6	0	Other	Healthy	3	N/A	Within-day	3
Franzinelli 1976 [217]	Cohort	FEV1	5	25.0		Other	Healthy	9	N/A	Within-day	3
Michoud 1982 [218]	Cohort	FEV1	5	26.0	68.2	Other	Healthy	10	N/A	Within-day	1
Williams 2015 [219]	Case-control	Both	8	26.0	100.0	Other	Healthy	3		Between-day	1
Cochrane 1977 [220]	Cohort	Both	10	26.6	50.0		Healthy	6	7 to 7	Between-day	1
Mccarthy 1975 [221]	Cohort	Both	20	27.0	58.3	Other	Healthy	10	7 to 7	Between-day	1
Koh 1997 [222]	Cohort	Both	22	28.3	91.-	Other	Healthy	3	1 to 1	Between-day	1
Randell 1999 [223]	Cohort	Both	11	28.5	36.3	Other	Healthy	16	1 to 7	Between-day	1
Gimeno 1993 [224]	Cohort	FEV1	7	30.5	85.7	Other	Healthy	6		Between-day	1
Malmberg 1993 [225]	Cohort	Both	10	32.0	10.0	Other	Healthy	8	N/A	Within-day	1
Zou 2013 [226]	Cohort	Both	22	32.3	91.0	Other	Healthy	2	1 to 1	Between-day	3
Chinn 1994 [227]	Cohort	Both	34	33.0	100.0	Primary	Healthy	2	N/A	Within-day	1
Lebowitz 1982 [228]	Cohort	Both	10	33.7	20.0	Other	Healthy	150			1

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex %male (primary analysis)	Setting	Health status	Number of measurements	Range of time between measurements (days)	Variation measurement interval	Risk of Bias score
Lam 1981 [229]	Cohort	Both	78	35.3		Primary	Healthy		N/A	Within-day	1
Lawther 1974 [230]	Cohort	Both	4	35.5	100.0	Primary	Healthy	837		Between-day	3
Struthers 1988 [231]	Cohort	Both	10	35.5		Other	Healthy	10	N/A	Within-day	1
Chinn 1977 [232]	Cohort	FEV1	24	36.1	54.2	Other	Healthy	10	N/A	Within-day	3
Neild 1989 [233]	Cohort	FEV1	25	36.5	44.0	Other	Healthy	3	1 to 7	Between-day	1
Humerfelt 1998 [234]	Cohort	Both	26368	38.0	100.0	Other	Healthy	3	N/A	Within-day	2
Rozas 1982 [235]	Cohort	Both	15	38.5	46.7	Other	Healthy	5	1 to 1	Between-day	3
Wang 2000 [236]	Cohort	FEV1	160	38.6	99.4	Primary	Healthy	11	180 to 1825	Between-day	1
Gulsvik 2001 [237]	Cohort	Both	480	39.2	40.0	Primary	Healthy	3	N/A	Within-day	1
Jensen 2007 [238]	Randomised crossover/cohort	FEV1	11	39.3	36.3	Other	Healthy	3	90 to 90	Between-day	2
Baur 1996 [239]	Cohort	FEV1	26	39.9	42.3	Other	Healthy	51	1 to 60	Between-day	2
Kainu 2008 [240]	Cohort	Both	603	48.2	41.1	Primary	Healthy	3	N/A	Within-day	1
Enright 1991 [241]	Cohort	FEV1	5887	48.5	62.9	Primary	Healthy	2	25 (mean)	Between-day	1
Hamedani 2013 [242]	Cohort	Both	8	54.0	50.0	Other	Healthy	3	3 to 5	Between-day	1
Beckett 1997 [243]	Cohort	FEV1	105		88.0	Primary	Healthy	4	365 to 365	Between-day	3
Biddiscombe 2009 [244]	Cohort	Both	24		50.0	Other	Healthy	18	1 to 1	Between-day	3

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex %male (primary analysis)	Setting	Health status	Number of measurements	Range of time between measurements (days)	Variation measurement interval	Risk of Bias score
Billings 2014 [245]	Cohort	FEV1					Healthy		365 to 365	Between-day	3
Burki 1976 [246]	Cohort	Both	5		60.0	Other	Healthy	6	1 to 5	Between-day	3
Garrard 1986 [247]	Cohort	Both	6		100.0	Other	Healthy	6			1
Houghton 2005 [248]	Cohort	FEV1	24			Other	Healthy	2	N/A	Within-day	3
Liu 2018 [249]	Cohort	Both	4		25.0	Other	Healthy				1
Wise 1995 [250]	RCT	Both	5885	48.9	62.9	Other	COPD	2	24.9 (mean)	Between-day	1
Rowan 2014 [251]	Cohort	FEV1	30	56.7	50.0	Other	COPD	2	14 to 14	Between-day	2
Holmner 2020 [252]	Cohort	FEV1	11	57.0	27.2	Other	COPD	2	N/A	Within-day	3
Taube 2003 [253]	Cohort	Both	13	61.0	76.9	Other	COPD	9	1 to 5	Between-day	1
Owens 1986 [254]	Cohort	Both	13	63.0	92.3	Other	COPD	2	60 to 60	Between-day	1
Anderson 2017 [255]	Cohort	Both	69	64.1	69.3	Other	COPD	2	14 to 42	Between-day	1
Mahler 2009 [256]	Cohort	Both	101	66.0	48.5	Other	COPD	2	3 to 7	Between-day	3
O' Donnell 1998 [257]	Cohort	Both	29	67.0	75.9	Other	COPD	4	14 to 14	Between-day	1
Nosedá 1989 [258]	Cohort	FEV1	20	67.8	100.0	Other	COPD	4	30 to 30	Between-day	1
Bofan 2013 [259]	Cohort	Both	24	68.0	79.2	Other	COPD	2	6 to 6	Between-day	1
Sirichana 2014 [260]	Cohort	FEV1	15	69.8	43.0	Other	COPD	30	1 to 1	Between-day	3

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex %male (primary analysis)	Setting	Health status	Number of measurements	Range of time between measurements (days)	Variation measurement interval	Risk of Bias score
Chan-Thim 2014 [261]	Cohort	Both	14	71.0	64.2	Other	COPD	4	1 to 7	Between-day	1
Laube 2019 [262]	Baseline measures of RCT	FEV1	9	72.4	66.6	Other	COPD	2	56 to 56	Between-day	1
Yamamoto 2021 [263]	Cohort	Both	85	72.4	91.8	Other	COPD	3	N/A	Within-day	1
Van Noord 1989 [264]	Cohort	Both	53	30.0	52.9	Other	Asthma	2	N/A	Within-day	1
Faul 1999 [265]	Cohort	FEV1	9	30.2	77.8		Asthma	3	14 to 56	Between-day	1
Huang 2021 [266]	Cohort	FEV1	12	41.1	67.0	Other	Asthma	55	1 to 28	Between-day	1
Palikhe 2019 [267]	Cohort	FEV1	19	43.0	36.8	Other	Asthma	4	90 to 90	Between-day	3
Jung 2011 [268]	Cohort	FEV1	45	46.6	48.9	Other	Asthma	2	7 to 7	Between-day	3
Roberto 2015 [269]	Cohort	Both	93	57.3	54.0	Other	Asthma	2	N/A	Within-day	3
Donaire 2015 [270]	Cohort	Both	93	57.4	54.0	Other	Asthma	3	N/A	Within-day	1
Alamer 2015 [271]	Cohort	FEV1	55				Asthma	3	N/A	Within-day	3
Baptist 2008 [272]	Cohort	FEV1	35		49.0	Other	Asthma	3	N/A	Within-day	1
Barley 2006 [273]	Cohort	FEV1	144			Other	Asthma	5	8 to 29	Between-day	3
Dennis 2011 [274]	Cohort	FEV1	94			Primary	Asthma	3	N/A	Within-day	3
Silkoff 2016 [275]	Cohort	Both	158				Asthma	4	90 to 90	Between-day	1
Stanbrook 2004 [276]	Cohort	FEV1	21	27.5	33.3	Other	CF	31	1 to 1	Between-day	1

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex %male (primary analysis)	Setting	Health status	Number of measurements	Range of time between measurements (days)	Variation measurement interval	Risk of Bias score
White 2017a [277]	Cohort	FEV1	106	29.8	43.4	Other	CF	6			1
White 2017b [278]	RCT	FEV1	13	29.8		Other	CF	2			1
Bayfield 2014 [279]	Cohort	FEV1	136				CF	2			3
Shaw 2014 [280]	Cohort	FEV1				Other	CF			Between-day	3
Lindgren 1997 [281]	Cohort	Both	77	47.9	40.2	Other	Other	8.7	1 to 1	Between-day	3
Kafaja 2018 [282]	Pooled data from both arms of two RCTs	FVC	300	50.3		Other	Other			Between-day	3
Sheshadri 2018 [283]	Cohort	Both	51	55.0	33.3	Other	Other	133	7 to 7	Between-day	2
Moor 2021 [284]	Cohort	FVC	10	60.3	40.0	Other	Other	42	1 to 7	Between-day	1
Tan 2017 [285]	Cohort	Both	33	62.0	55.0	Other	Other	3	N/A	Within-day	3
Krug 2011 [286]	Baseline measures of RCT	FVC	216	64.0	84.0	Other	Other	2	3 to 16	Between-day	2
Berry 1974 [287]	Cohort	FEV1	28		100	Primary	Other	36	1 to 1095	Between-day	2
Glindmeyer 1978 [288]	Cohort	Both	9	31.0	55.6	Other	Multiple	10	1 to 1	Between-day	1
Fonseca 2005 [289]	Cohort	FEV1	38	33.0	29.0	Other	Multiple	3	N/A	Within-day	3
Moore 2009 [290]	Cohort	FEV1	13	39.0		Other	Multiple	3.5	N/A	Within-day	1
Klarholt 1979 [291]	Cohort	FEV1	80	45.0		Other	Multiple	3	N/A	Within-day	2

Paper	Study type	FEV1, FVC or both	Number of individuals	Age (primary analysis)	Sex %male (primary analysis)	Setting	Health status	Number of measurements	Range of time between measurements (days)	Variation measurement interval	Risk of Bias score
De Vries 2018 [292]	Cohort	FEV1	24				Multiple	33			3
Rossmann 2011 [293]	Cohort	Both	8			Other	Multiple	3	5 to 7	Between-day	3
Median			44.5	45.0	43.4			4			
Total			42362								

Setting Primary=primary or community; Other=tertiary/secondary/laboratory. Risk of bias are presented as highest score, with questions 4 and 5 excluded. 1=very good; 2=adequate; 3=doubtful

With questions four and five of the risk of bias scoring excluded, 56 (53.8%) studies scored the best risk of bias score of one, 10 (9.6%) studies scored a risk of bias of two, and 38 (36.5%) studies scored a risk of bias of three. Risk of bias scores are presented in tables 4.1 and 4.2.

Results of individual studies

The following summary statistics refer to the primary population as defined above. One hundred and one (97.1%) studies reported the number of individuals in the study (range 4 to 26,368); 87 (83.7%) reported age (range 0.1 to 72.4 years); 99 (95.2%) reported the number of measurements per individual (range 2 to 837); 82 (78.8%) reported sex (range 0% to 100% male). Of the 104 studies, 19 (18.3%) were available only as abstract, the rest had full texts available. Only 12 studies (11.5%) explicitly recorded ethnicity. Sixty-six (63.5%) studies reported a CV, 14 (13.5%) reported a SD and 39 (37.5%) reported an ICC.

All studies were cohort design except for six which were the baseline or placebo group of a randomised controlled trial, and one case-control trial. Of the 97 studies reporting study setting 21 (20.2%) were conducted in primary care and 76 (73.1%) studies were conducted in hospitals, laboratories and other secondary or tertiary care setting. In seven studies the setting was unclear.

Aims of the studies included explicitly measuring variation in healthy populations and populations with disease, assessing the accuracy of spirometry analytical equipment and use of variation to assess self-reported adherence.

Forty-five (43.3%) studies used healthy populations, 19 (18.3%) were on patients with asthma, 14 (13.5%) were on patients with COPD, 8 (7.7%) were on patients with cystic fibrosis, 7 (6.7%) were on multiple pooled health statuses and 11

(10.6%) studies were on patients with other health conditions e.g. mesothelioma or post-lung transplant.

Fifty-one studies (49.0%) reported in units of Litres, 8 (7.7%) reported in units of mLs, 1 (1.0%) in units of Litres squared/year, and 24 (23.1%) in units of % predicted. In 20 (19.2%) studies the units were missing or unclear. Sixty-one studies (58.7%) reported between-day variation and 36 (34.6%) reported within-day variation. In 7 studies (6.7%) this information was missing or unclear.

The median CV (within-individual coefficient of variation) of FEV1 was 0.034 (IQR 0.025 to 0.067) and for FVC was 0.031 (0.020 to 0.048). Figure 4.2 and figure 4.3 show a forest plot of the within-individual CVs of FEV1 and FVC respectively, for all the studies reporting results in units of litres, split by health status. For healthy individuals the CV of FEV1 ranged from 0.011 to 0.177 with a median of 0.028, and for FVC the CV ranged from 0.010 to 0.163 with a median of 0.029. For individuals with COPD the CV of FEV1 ranged from 0.020 to 0.122 with a median of 0.068, and for FVC the CV ranged from 0.016 to 0.140 with a median of 0.068. For individuals with asthma, the CV of FEV1 ranged from 0.018 to 0.089 with a median of 0.030 and for FVC the CV ranged from 0.014 to 0.029 with a median of 0.024. It can be seen that FEV1 and FVC are more variable over time in individuals with COPD than in healthy individuals. The number of studies with individuals with asthma is too small to draw a firm conclusion regarding variability in this subgroup, although it appears to be smaller than individuals with COPD.

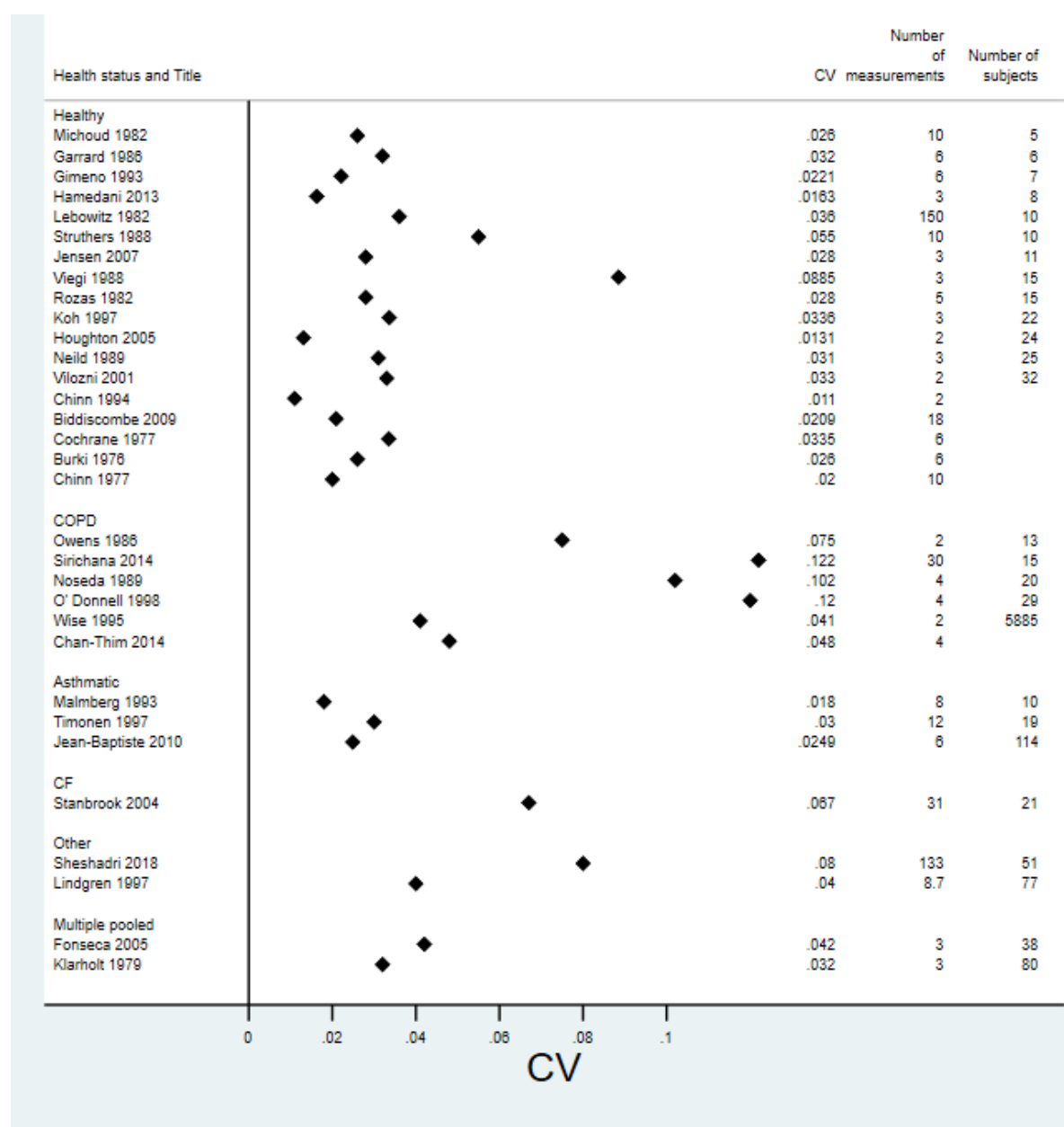


Figure 4.2 Forest plot of within-individual CV of FEV1 by health status, all studies using units of Litres.

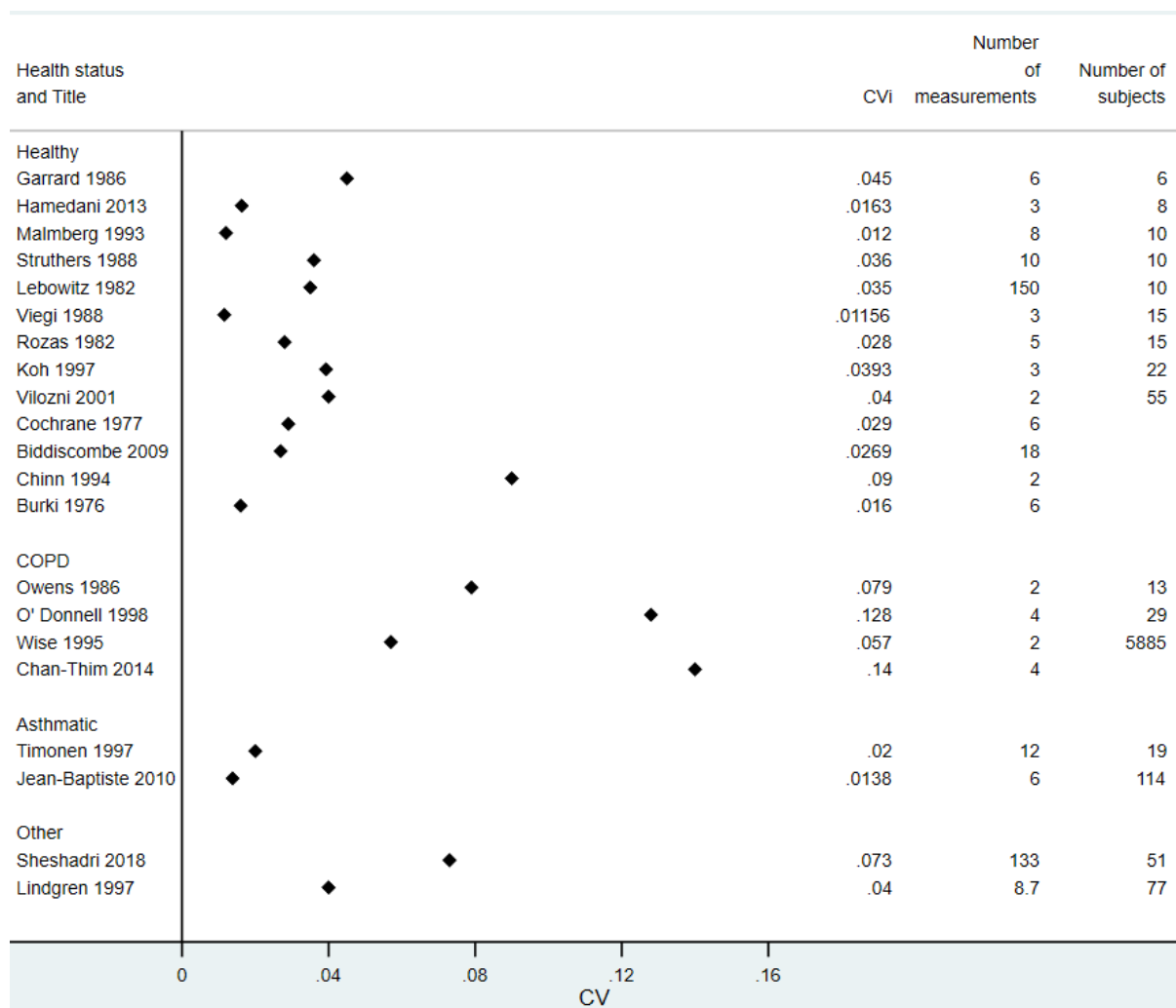


Figure 4.3 Forest plot of within-individual CV of FVC by health status, all studies using units of Litres.

Figures 14.1 and 14.2 in Appendix 4.3 show forest plots of CVs of FEV1 and FVC split by setting, but there are too few studies in primary care settings to draw conclusions. Figures 14.3 and 14.4 in Appendix 4.3 illustrate the distribution of CVs split by between or within-day intervals of measurement using units of Litres. The within-day CVs for FEV1 reported in this review for studies using units of L ranged from 0.011 to 0.089 with a median of 0.026 and for FVC from 0.012 to 0.090 with a median of 0.025. This compares with between-day CVs for FEV1 ranging from 0.016 to 0.120 with a median of 0.034 and for FVC 0.016 to 0.140 with a median of 0.039. These show that the variation of between-day measurements is subjectively higher than the within-day measurements.

Figures 14.5 and 14.6 show FEV1 and FVC CVs split by unit of measurement (Litres, mLs or % predicted), with no obvious difference between these groups, although the numbers of studies for % predicted and mL were low, so this lack of difference should be interpreted cautiously. However, figures 14.7 and 14.8 which show FEV1 and FVC CVs by risk of bias suggest a higher variation in studies with a lower risk of bias. Figures 14.9 and 14.10 present forest plots of SDs of FEV1 and FVC split by health status for those studies using units of litres.

Figures 4.4 and 4.5 show forest plots and meta-analyses of the ICCs of FEV1 and FVC in studies showing units of litres split by health status. The pooled estimate for ICC for FEV1 of all individuals was 0.93 (95% CI 0.92 to 0.95) and for FVC was 0.95 (95% CI 0.93 to 0.96). For healthy individuals, ICCs of FEV1 ranged from 0.82 to 1.00 with a median of 0.98, and ICCs of FVC ranged from 0.85 to 1.00 with a median of 0.96. These results suggest excellent reliability of the tests.

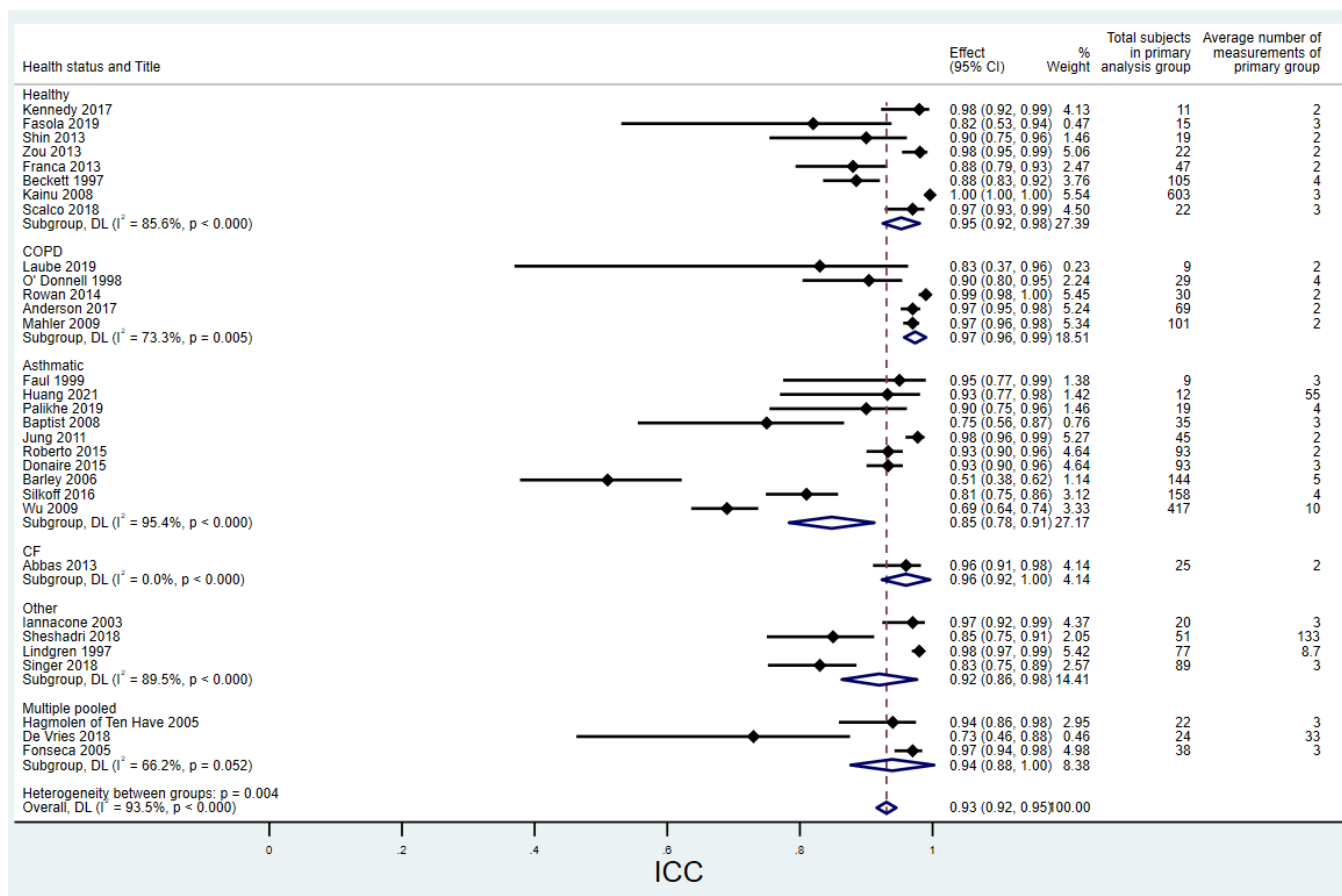


Figure 4.4 Forest plot and meta-analysis of ICCs of FEV1 by health status, all studies using units of Litres.

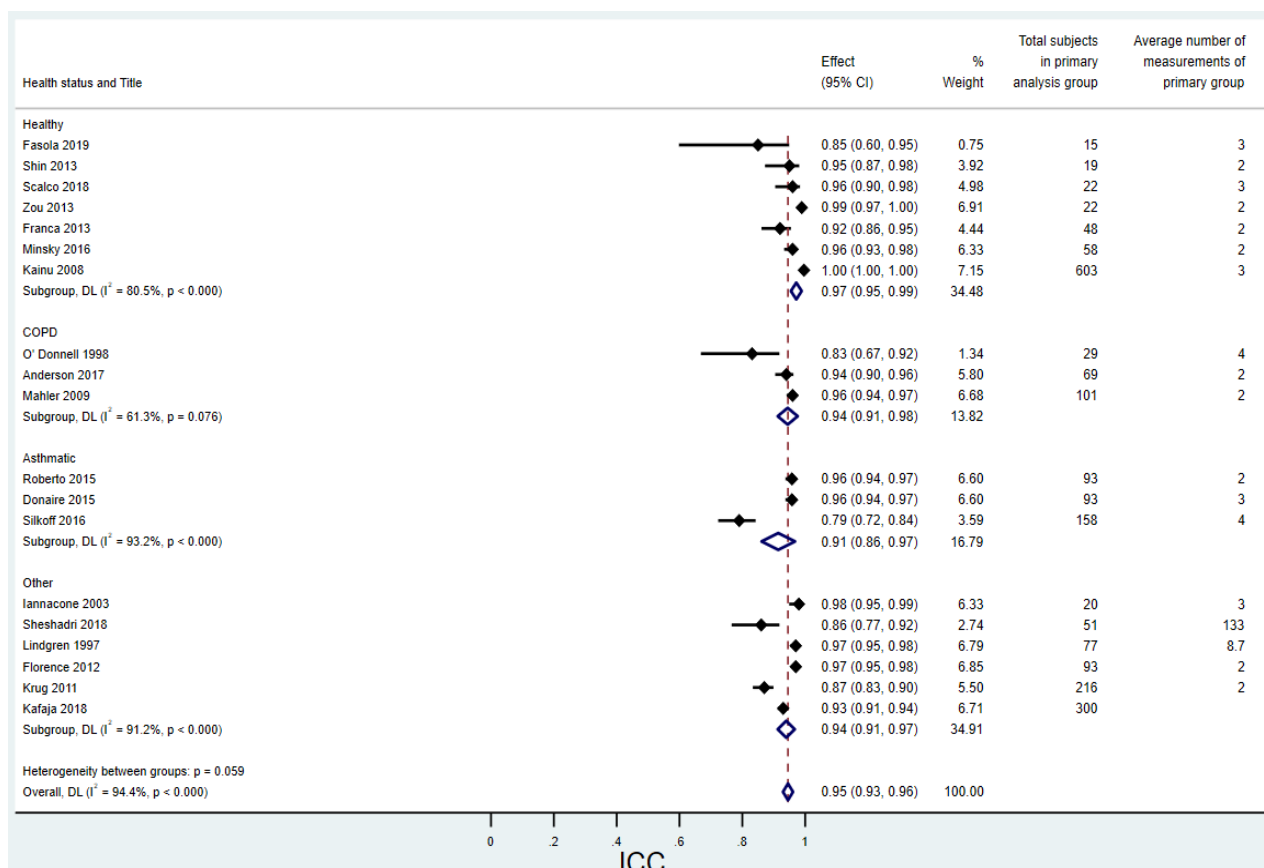


Figure 4.5 Forest plot and meta-analysis of ICCs of FVC by health status

4.6 Discussion

This is the first systematic review of within-individual variation of the spirometry measures FEV1 and FVC. The aim was to describe the variation of spirometry over time within an individual, and the results describe the degree of variation between days. The reliability of the tests appeared to be excellent when looking at ICC estimates, but the results of the CV estimates suggests there may be enough within-individual variation to be clinically important.

The between-day variation in this review is subjectively a little higher than the within-day variation, but the difference isn't great, suggesting that within-day variation (i.e. analytical and pre-analytical variation) and between-day variation both contribute to the overall real-world variation in FEV1 and FVC measurements seen

by clinicians.

This analysis indicated that individuals with COPD have higher variation than healthy individuals, with median CVs in healthy individuals for FEV1 and FVC respectively of 0.028 and 0.029, compared with the median CVs in individuals with COPD for FEV1 and FVC respectively of 0.068 and 0.068. However, sample sizes were generally smaller in studies involving COPD. The paper that had over five thousand participants did have a CV for FEV1 of 0.041 which was lower than the other COPD studies, but this was still higher than most of the studies on healthy individuals. Interestingly, the same did not hold true for individuals with asthma, despite the fact that asthma is considered to have highly variable symptoms, with median CVs in individuals with asthma for FEV1 and FVC respectively of 0.037 and 0.024 [294].

Little difference in subgroup analysis was found when the studies were split by setting or unit of measurement, see Figures 14.6, 14.7, 14.10 and 14.11 in Appendix 4. Similarly risk of bias seems to have had little impact on the results, see figures 14.12 and 14.13 in Appendix 4.

The patient populations of the included studies were diverse in terms of age, ethnicity, sex and health status. Many of the studies provided analysis of the variation of the outcome measures in subgroups. To choose which group to include in the meta-analysis was problematic, since it had to involve arbitrary choices. However, the rules to decide which group to include were applied systematically as described in the methods. There were missing values in a number of studies for age, sex and ethnicity, with ethnicity being particularly poorly recorded. Some studies reported demographic data for the whole population but not for the subgroup for which variation data was reported. In these cases, the demographic data for the

whole study population was used and was assumed to be close to the variation subgroup. These problems illustrate a high degree of heterogeneity in the included studies.

High quality studies with large numbers of individuals specifically designed to test the within-individual variation of spirometry are lacking. If a spirometry test has a high within-individual variation, this has implications for clinical decisions. For example, a value that is abnormal due to random chance because of a high within-individual variation might lead to an inaccurate positive or negative diagnosis, or an incorrect decision on whether to begin or withhold medication. It has been stated that a change from baseline of 5-10% is considered clinically important [295], which suggests that the variation demonstrated in this review could have important implications for diagnosis and monitoring. Variation should also be taken into account when clinical guidelines are being devised. Studies are therefore needed that more accurately demonstrate the within-individual variation of spirometry tests over time.

Studies on spirometry measurements with multiple measures over longer periods of time would demonstrate the total within-individual variation of these tests, encompassing the measurement error and the biological variation. This total variation is the most useful measure for clinicians since it provides guidance on the “real-world” variation of the selected tests, and hence the best idea of whether a single measurement can be relied upon, and if not, how many measurements are required to make meaningful clinical decisions. However, decline in lung function must be distinguished from long-term variation in any future studies. A study into the extent of how the variation in spirometry values changes compared with time between measures would help address the knowledge gap of how variation of

spirometry changes over time.

Limitations

There are a few limitations with this review. One is the possibility of relevant studies being missed by the search strategy. However, the search strategy was as broad as possible, and a large number of abstracts and studies were screened. On the other hand, this broad search strategy could have led to the inclusion of poor-quality studies. Note this systematic review is only concerned with within-individual variation, so some large cohort studies of COPD and asthma which document decline were excluded since they provided no information on variation.

Ideally, 100% of the data extraction would have been checked by a second reviewer due to the risk of error when transcribing complex numerical data. 50% verification was chosen for pragmatic reasons due to time limitations of the second reviewer. However, the level of agreement for the 50% checked was subjectively very high.

Risk of bias due to missing results was deemed to be low, since in most cases the measure of variation was not the primary outcome of the individual studies, and therefore systematically withholding publication based on measurement variation is unlikely. Some studies with repeated measures do not report variation data (and so are excluded from this review), but this is likely to be because variation is not relevant to the study's primary outcome, and therefore is unlikely to introduce a systematic bias.

Limitations with the risk of bias tools are discussed in Chapter 2.6.

Note that some studies may not have clearly distinguished between a material change in the disease in populations that have the disease (e.g. an increase in mean value) and an increased level of variation due to the disease (the two aspects of pathological variation), which could increase the measured variability and influence the association of increased variability with presence of disease/increased disease severity.

Note also that some of the underlying diseases for which tests are performed have a higher variability than others, that is pathological variation will comprise a higher proportion of the total variability.

There is no well-described, validated method for meta-analysis of CVs therefore it was not undertaken in this study.

4.7 Conclusion

This comprehensive systematic review shows that the published data for variation of spirometry measurements is of limited amount, quality and generalisability.

Having shown there is limited quality data on within-individual measured variation in spirometry, in the next chapter I report the findings of a systematic review of depression symptom measurement instruments. These are a form of patient reported outcome measure (PROM), which are commonly used in general practice to aid in clinical decision-making, particularly for symptoms that are hard to measure objectively such as pain and mental health. These instruments perform a different type of measurement from laboratory tests and from those measured by medical staff such as spirometry, because they involve a self-reported questionnaire.

5 Chapter five: Within-individual Variation of Measured Depression Symptoms: A systematic review and meta-analysis

A version of this chapter has been previously published as:

Gough, A, Marshall, T, Ferris, E & Sitch, A, Within-individual variation of measured depression symptoms: A systematic review and meta-analysis, Journal of Affective Disorders Reports, Volume 14, 2023, 100675, ISSN 2666-9153, <https://doi.org/10.1016/j.jadr.2023.100675>.

5.1 Abstract

Background

Depression screening instruments are commonly used to assess the presence and severity of depression symptoms. However, there is little information on variation in depression screening scores over time within the same individual.

Methods

A systematic review and meta-analysis was performed of studies reporting the within-individual variation of the Beck's Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS), Patient Health Questionnaire-9 (PHQ-9) and Patient Health Questionnaire-2 (PHQ-2). Multiple databases were searched from inception

to 14th July 2023. Title and abstract screening was performed in duplicate, full text screening and data extraction by one reviewer and verified by a second. Risk of bias was assessed with a modified COSMIN tool.

Results

Of 2,798 titles and abstracts and 157 full text articles screened, 41 met the inclusion criteria. No studies were on patients with depression, and most had only two measurements, less than two weeks apart. The pooled estimates of ICCs (Intraclass Correlations) and 95% confidence intervals for BDI, HADS, PHQ-9 and PHQ-2 were 0.89 (0.84-0.93), 0.89 (0.85-0.93), 0.84 (0.81-0.88) and 0.75 (0.56-0.93), respectively. Studies in healthy subgroups showed lower ICC than those with physical illness.

Limitations

Assessment of variation was not the main aim of most studies assessing depression measurement instruments, so it is possible that some relevant studies have been missed.

Conclusions

Within-individual scores for BDI, HADS and PHQ-9, show generally excellent agreement over periods of up to two weeks. However published data on within-individual variation is lacking over longer time periods and for patients with depression.

5.2 Aim of review

To describe and evaluate the current literature on the within-individual variation of selected depression symptom measurement instruments in healthy subjects and patients with medical conditions and to estimate within-individual variation of these measures.

5.3 Introduction

Depression is a mental disorder characterised by loss of positive affect, low mood and behavioural, cognitive, emotional and physical symptoms [296]. Since there is no objective diagnostic test for depression the “gold standard” diagnosis is a structured clinical interview [297-299]. Unfortunately, this takes around ninety minutes to complete and requires a large degree of clinician training [300].

However, measurement-based care is becoming standard practice and involves ongoing collection of outcome measures including depression measures. Depression measurement instruments have been developed for use in non-psychiatric settings to screen for depression symptoms, to assess severity and for monitoring (e.g. response to treatment). These are less time-consuming and do not require clinician training.

Validated depression measurement instruments, in particular the PHQ-9, BDI and HADS were recommended by NICE. Their use was incentivised under the Quality Outcomes Framework (QOF) for both initial management and monitoring of depression in the UK between 2006 and 2013 [301]. Current NICE guidelines still

recommend using a validated measure to inform and evaluate treatment, but this is no longer incentivised under QOF [302].

Problems with diagnosis, assessment of severity and monitoring of depression symptoms arise due to factors such as the imperfect sensitivity and specificity of depression measurement instruments, practitioners' concerns about the reliability and practicality of depression measurement instruments, and problems with reliance on clinical judgement. Another problem is biological variation, the natural variation within an individual of the severity of the condition over time. This is partly related to factors such as season, diurnal variation and reproductive cycles, disease progression or a response to treatment [303]. However, most within-individual biological variation is chance variation [15]. There is biological variation in mood and measurement variation in its assessment, which lead to within-individual variation of measured depression symptoms. Because mental health measurement instruments produce categorical scores, within-individual variation can impact the interpretation of measurement findings. This has the potential to lead to inaccurate conclusions about symptom progress and treatment outcomes[304] [305].

The depression measurement instruments most commonly used when screening for depression in a non-psychiatric setting [297] are: Patient Health Questionnaire-9 (PHQ-9); Patient Health Questionnaire-2 (PHQ-2); Hospital Anxiety and Depression Scale (HADS); Beck Depression Inventory first and second editions (BDI and BDI-II). This review was therefore restricted to these measurement instruments, and other measurement instruments were not included in order that the scope of the review not be too broad. Details on these instruments such as structure, scoring method and sensitivity/specificity can be found in Appendix 5.1.

The problem of within-individual variation in measurement using these instruments has received little consideration. An unpublished scoping review (see Appendix 10.1, 20.1.4) ,identified only one paper on within-individual variation of measurement scales for depression and no systematic review.

The purpose of this review therefore is to review the literature on within-individual variation of commonly used depression measurement instruments and estimate the extent of within-individual variation in measured depression symptoms.

5.4 Methods

PRISMA and PRISMA-S guidelines were followed in the reporting of this review [134]. The methods for this systematic review were as described in Chapter 2 for the HbA1c systematic review, except as noted below.

Searches were devised to identify studies with multiple depression scores measured within the same individual. These included cohort studies, clinical trials or any studies in which a depression measurement was repeated more than once in the same individual.

Any study was included if it recorded primary research data on the variation of at least two scores from the same depression measurement instrument (PHQ-9, PHQ-2, BDI or HADS) within the same individual.

The protocol was registered at Prospero registration number CRD42020213398.

The review protocol is available at

https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020213398.

Medline, Embase, APA Psycinfo, Cochrane Central, Epistemonikos and Open Grey were searched from inception to 9th October 2020 and an updated search was performed up to 14th July 2023 (full details in Appendix 5.2). Search terms were adapted for each database searched. Subject experts were contacted for suggestions for further studies. The references of included studies were checked by hand for further relevant studies. PROSPERO was checked for ongoing reviews, and the protocol was registered with PROSPERO (Ref CRD42020213398). See Appendix 5.3 for sample search strategy used in searching Medline and Embase.

Table 5.1 lists the outcome and other main variables extracted. Variables selected for extraction and for subgroup analysis were those I hypothesised to most likely affect variability such as age and disease status. All eligible outcomes (measures of variation for PHQ-9, PHQ-2, BDI, HADS) were included.

Table 5.1 Variables extracted from studies

Variable	Definition and rules
Author name and year of publication	Full author list extracted. Short title given as first author and year of publication
Depression measurement instrument	Depression measurement instrument reported in paper: PHQ-9, PHQ-2, BDI, HADS (or multiple)
Study design	Cohort, RCT etc
Number of subjects	Number of subjects used to calculate variation measure
Age	Average age of subjects used to calculate variation measure. If the subjects used to calculate variation measure is a subset of the study population and only the age is given for the whole population, the age of the whole population is used.
Sex	Percentage of subjects used to calculate variation measure who were male. If the subjects used to calculate variation measure is a subset of the study population and only the sex is given for the whole population, the sex of the whole population is used.
Ethnicity	Ethnicity recorded in study.
Setting	Primary care/community or secondary/tertiary/laboratory setting
Health status	Healthy/depressed/other mental health condition/physical health condition/mixture
Number of measurements	Number of repeated applications of depression measurement instrument to the same individual
Time interval between measurements	Length of time between measurements in days
CV	Coefficient of variation of repeated measures within an individual
SD	Standard deviation of repeated measures within an individual
ICC	Intraclass correlation of repeated measures within an individual

The primary outcome measure was variation measure of repeated measurements within the same individual. Studies were pooled to provide combined results where possible (i.e. the same depression measurement instrument and the same measure of variation). If multiple depression measurement instruments were included, the data was extracted separately for each instrument.

Subgroup analyses were only performed on the PHQ-9 depression measurement instrument since it had the largest number of included studies. Where studies reported demographic data for the whole population but not for the subgroup for which variation data was reported, the demographic data for the whole study population was used and was assumed to be close to the variation subgroup.

5.5 Results

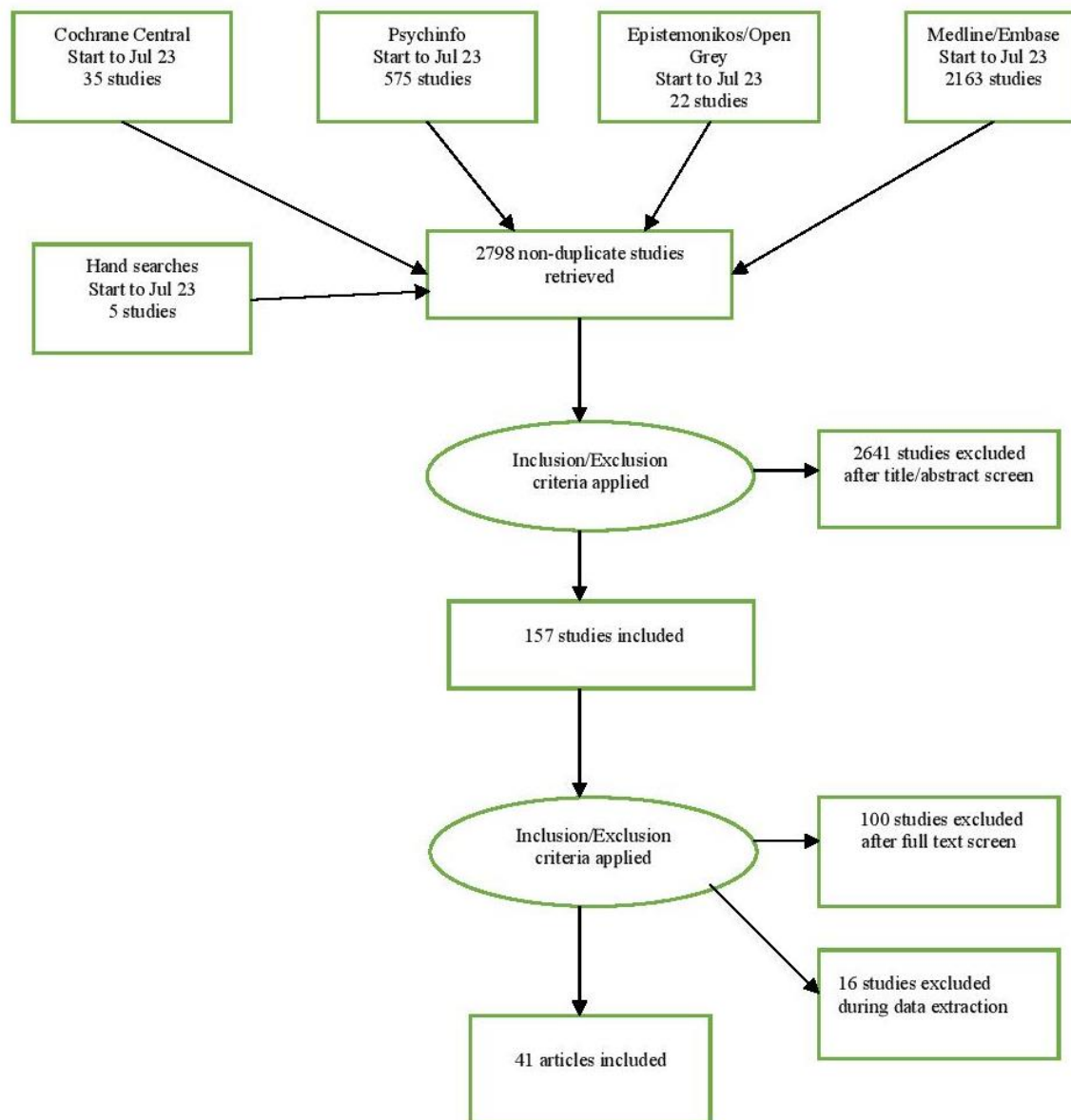


Figure 5.1: PRISMA flow diagram describing selection of studies.

Two thousand seven hundred and ninety-eight non-duplicate citations were retrieved after database searches of Embase, Medline, Psychinfo, Cochrane,

Epistemonokis and hand searching of reference lists of included studies. 2,641 articles were excluded after title and abstract screening leaving 157 articles included. After full text screening, 100 articles were excluded, and a further 16 were excluded during data extraction, leaving 41 included articles for analysis (see fig 5.1). Most full text exclusions were due to lack of reporting of a recognised measure of variation with the following exceptions: two studies that had duplicate populations to other included papers (the two duplicates were removed) ; one study not in with patients not in a steady disease state; one study involving a test that had a similar name but was different from the test of interest; one retracted article

Forty-one studies met the inclusion criteria, including between 15 and 458 subjects. Study populations were diverse in terms of age, gender and health status. Only two studies reported more than two repeat measures of the tests within the same subjects. The majority had time intervals between tests of one to two weeks. Ten studies included healthy populations or subpopulations. No studies looked at patients with depression as the main population or subpopulation. See Appendix 5.4 for details.

Risk of bias due to missing results was deemed to be low, since in most cases the measure of variation was not the primary outcome of the individual studies, and therefore systematically withholding publication based on measurement variation is unlikely. Some studies with repeated measures do not report variation data, because it is not of relevance to the study's primary outcome, but this is unlikely to introduce a systematic bias.

No study scored more than doubtful for question B4 *knowledge of previous scores when administering the test* and B5, *when assigning the scores* (questions B4

and B5). Appendix 5.4 presents the maximum risk of bias score for each study with questions B4 and B5 excluded, since every study would be rated at a high risk of bias with these scores included. Eight studies scored the lowest risk of bias, very good. Fifteen were ranked adequate and 18 were ranked doubtful. No studies were ranked inadequate, since these should have been excluded at the screening stage, for example due to an unstable population or lack of calculation of a recognised measure of variation. Appendix 5.5 lists the full risk of bias scoring for each study.

Results of individual studies

The following summary statistics refer to the primary population as defined in Synthesis Methods above, but include the studies by Karmisholt [306] and Shelton [307] which were excluded from the pooled results. All 41 studies reported number of subjects (range 15 to 458), age (range 15.9 to 71.4), number of measurements (range 2 to 13); 38 reported gender (range 0% to 100% male); and 39 reported time between measurements (range 1 to 84 days). Two out of 41 studies were available only as abstract, the rest had full texts available. Fifteen of the 41 studies explicitly recorded ethnicity. Forty studies reported an ICC, with results ranging from 0.232 to 0.980. One paper [306] reported a CV of 0.416 in patients with subclinical hypothyroidism. Table 5.2 shows the number of specific depression measurement instruments which were studied (53 depression measurement instruments in 41 studies). For the five studies where an ICC was reported and the time interval between measurements was 28 days or more, the ICC of the primary measurement ranged from 0.66 to 0.98. The number of measurements ranged from 2 to 13 but only three studies had more than two measurements.

Table 5.2 Depression measurement instruments analysed for variation in the included studies.

(total 53 groups or subgroups in 41 studies).

Depression measurement instrument	Number of studies
BDI-I	2
BDI-II	5
HADS	9
HADS - depression subscale only	5
PHQ-2	7
PHQ-9	25

All studies were cohort design except for two which used the patients from one arm of a randomised controlled trial (Weobong 2009 [308] and Lowe 2004 [309]). 11/41 studies were conducted in primary care (26.8%), 28/41 (68.3%) studies were conducted in hospitals, laboratories and other secondary or tertiary care settings and in 2/41 studies the setting was unclear (4.9%). In 30/40 (75%) studies that reported ICCs, the method of calculating the ICCs was not reported.

Almost all of the studies aimed to assess the reliability and repeatability of the screening instruments in a variety of settings, populations and health conditions.

Table 5.3 Health status of patients included in studies

Health status	Number of results (/41)	%
Healthy	12	29.3
Mental health condition	1	2.4
Physical health condition	28	68.3
Mix of mental and physical conditions	2	4.9

Table 5.3 shows the health status of the patients included in the 41 studies. Note that since the categories are not mutually exclusive, the % adds up to more

than 100. None of the studies identified expressly concerned individuals with depression.

Two studies were excluded from the pooled estimates, Karmisholt [306] because it reported a CV not an ICC and Shelton [307] because it reported a negative confidence interval which is mathematically impossible for an ICC.

The tau-squared result for the meta-analysis of ICCs of the four depression measurement instruments ranged from 0.001 to 0.006, suggesting a low between-study heterogeneity. There were few outliers, and the 95% confidence intervals of the weighted average results for the ICC are small. Appendix 5.6 shows the tau-squared results and predictive intervals for subgroup analyses.

On analysis grouped by depression measurement instrument, pooled estimates (95% CI) of ICC were similar for BDI, HADS, PHQ-9: 0.89 (0.84-0.93), 0.89 (0.85-0.93), 0.84 (0.81-0.88) respectively. For PHQ-2, ICC was 0.75 (0.56-0.93), see figure 5.2.

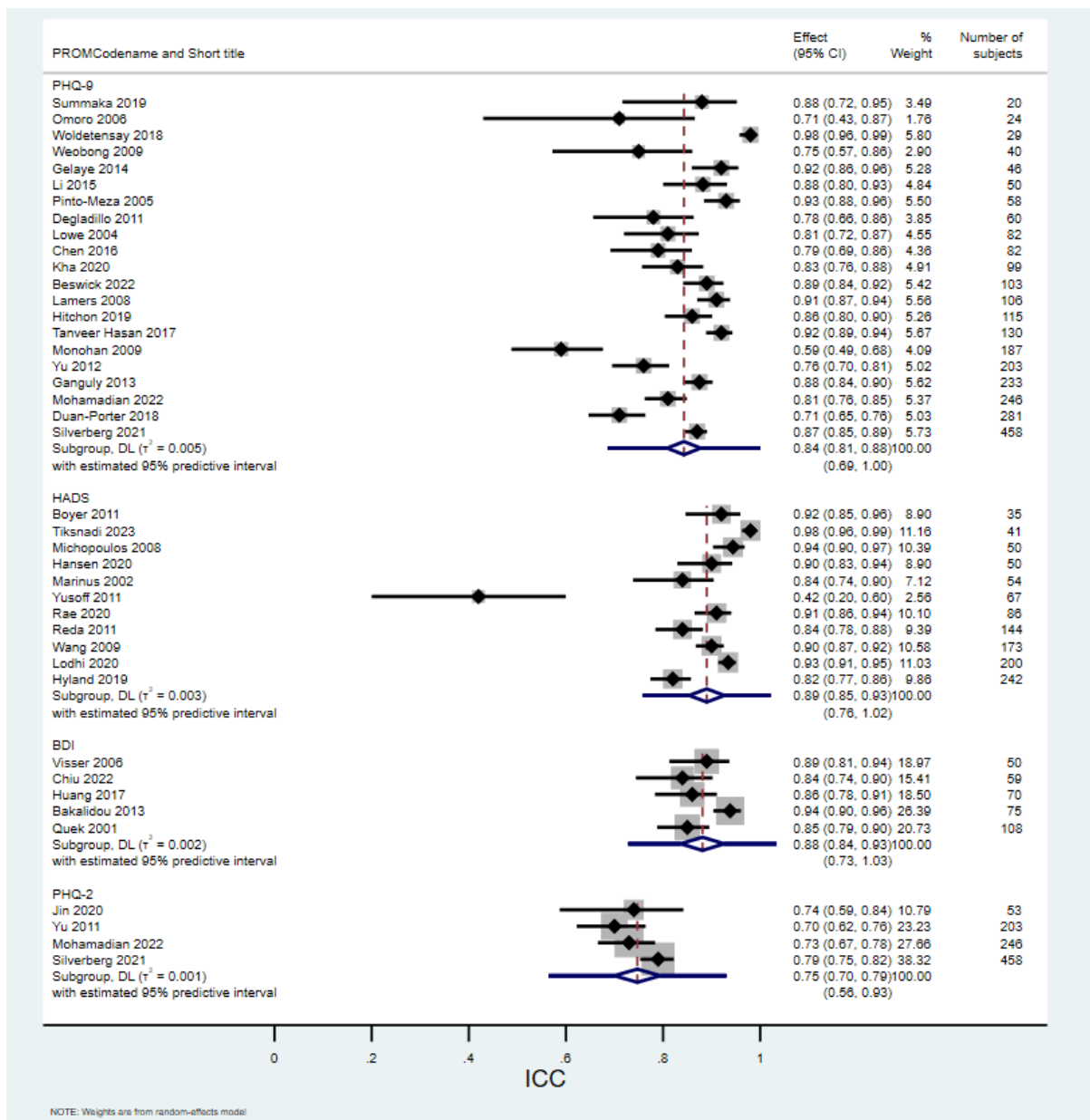


Figure 5.2 Forest plot of ICCs with 95% CI, grouped by depression symptom measurement instrument

There was little difference between the pooled estimates for primary care/community versus secondary care or other care with ICCs and 95% CIs of 0.86 (0.80-0.92) and 0.83 (0.79-0.87) respectively (figure 5.3).

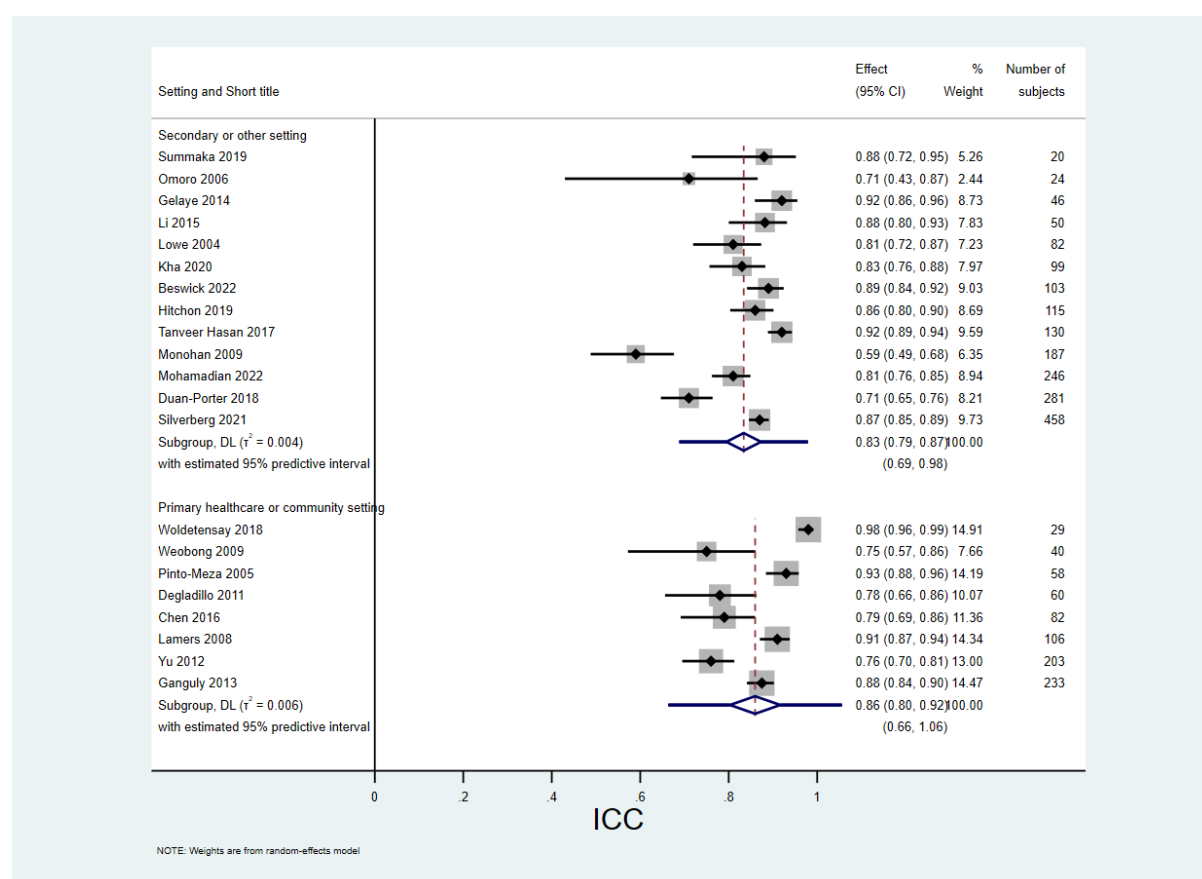


Figure 5.3 Forest plot of ICCs with 95% CI grouped by setting, PHQ-9 only

The pooled estimate of ICC in the healthy subgroup was lower, 0.79 (95% CI 0.73-0.85) than in the physical illness subgroup, ICC 0.88 (95% CI 0.84-0.91) although the confidence intervals overlapped (figure 5.4).

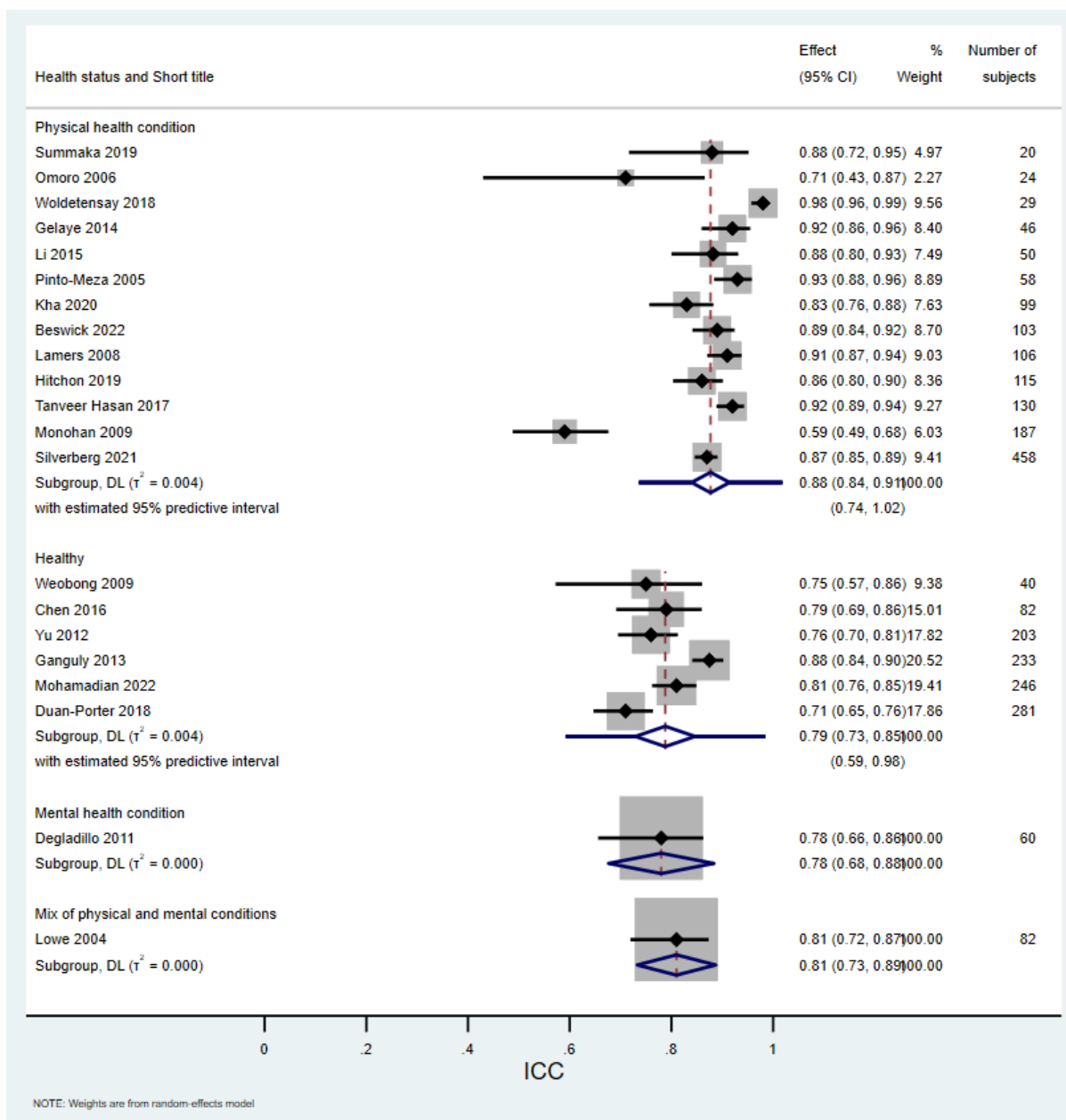


Figure 5.4 Forest plot of ICCs with 95% CI by health status, PHQ-9 only

When studies of 100 or less subjects were excluded (figure 5.5), the pooled ICC estimate for PHQ-9 was 0.83 (95% CI 0.80-0.86).

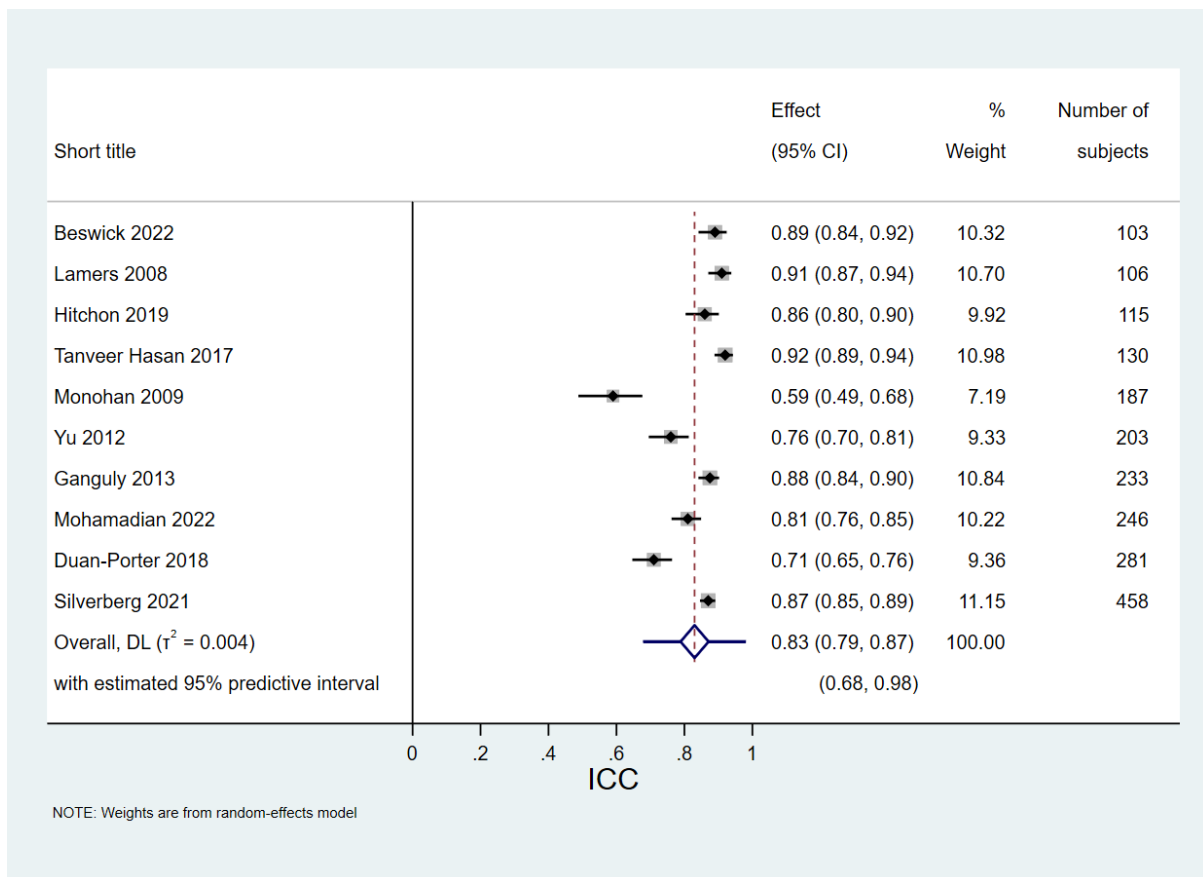


Figure 5.5 Forest plot of ICCs with 95% CI for those studies with more than 100 subjects, PHQ-9 only

Studies in which measurements were taken more than 7 days apart (figure 5.6) had a pooled estimate for PHQ-9 of 0.83 (95%CI 0.80-0.86).

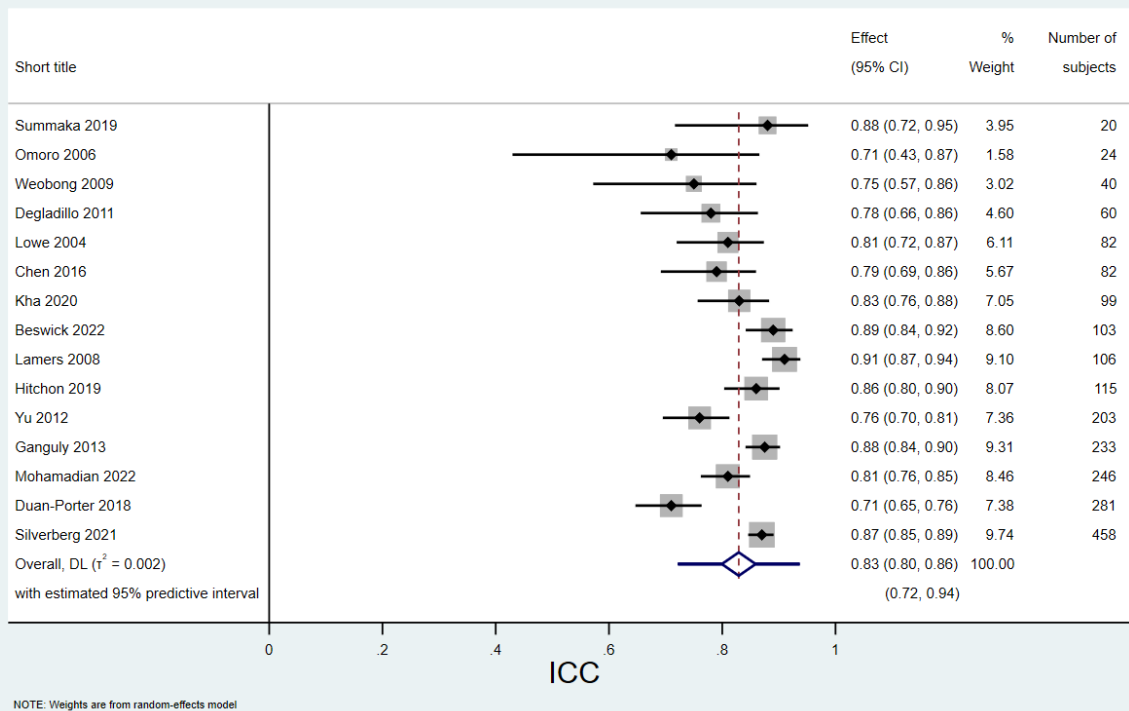


Figure 5.6 Forest plot of ICCs with 95% CI for those studies where measurements were performed more than 7 days apart, PHQ-9 only

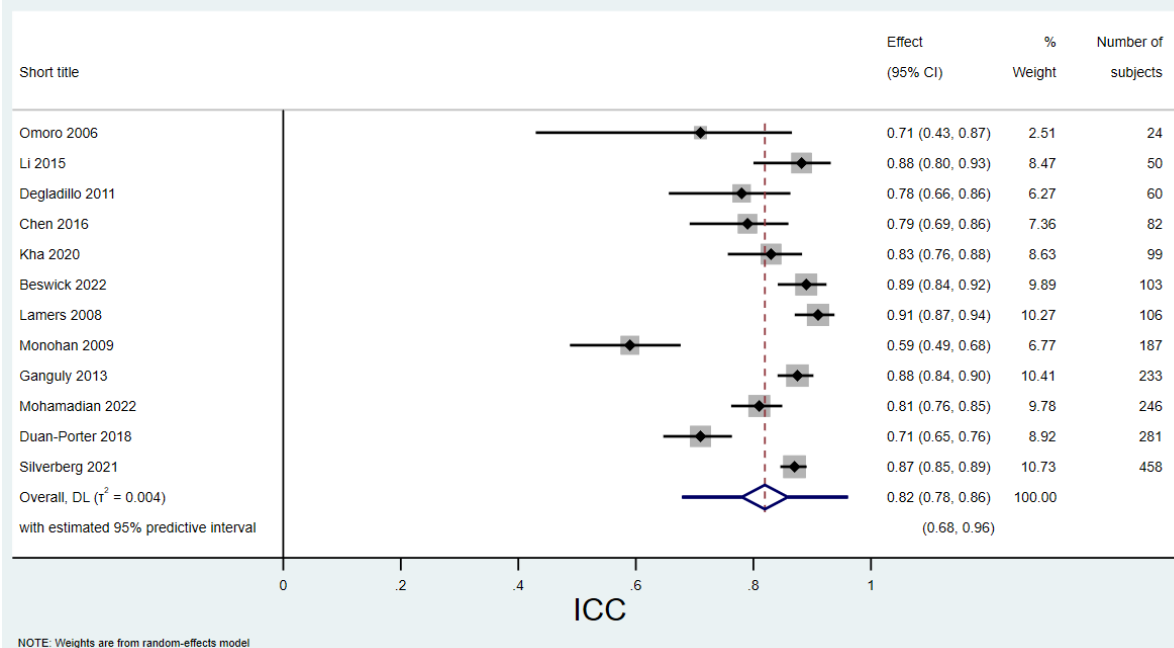


Figure 5.7 Forest plot of ICCs with 95% CI for those studies with a risk of bias of 2 or less, PHQ-9 only

When these results are compared to the full dataset where the pooled ICC estimate for PHQ-9 was 0.84 (0.81-0.88) it can be seen that there is little effect of sample size, time between measurements and risk of bias on the robustness of the result.

5.6 Discussion

This review describes the current literature on variation of scores derived from depression measurement instruments within an individual. The intraclass correlation (ICC) has been used in all but one of the included studies here to measure test-test reliability or variation. The pooled estimate of ICCs ranged from 0.84 to 0.89 for three of the depression measurement instruments (PHQ-9, BDI and HADS), suggesting generally these depression measurement instruments have excellent reliability, and the short-term variation is low. The healthy subgroup had higher variation than those with a physical illness, and the primary or community setting subgroup had lower variation than subgroups in other settings, though it is not clear why this is the case. Most studies used ICCs as the measure of reliability/variation, but one paper [306] used coefficient of variation (CV), and reported a markedly high variation (CV=0.416) which contrasts with the high reliability/repeatability shown by the other variation estimates.

The main aims of the included studies were to validate the selected tests. They were therefore not designed to measure variation. Sample sizes were small by general medical study standards, with the largest study population numbering only 458. In only three studies identified were more than two repeated measures performed. Two of these studies showed higher variation than most of the studies

with only two measurements. Duan-Porter [310] had 12 measurements one month apart, which resulted in an ICC for PHQ-9 of 0.71 which is lower than the pooled average. Karmisholt [306] had 13 measurements of HADS one month apart, which resulted in a high estimate of variation. By contrast, Beswick [311] had a lower variation, but only had three measurements performed over four weeks. The studies with more measurements over a longer time period had higher variation than most of the studies with only two measurements.

In only seven studies identified were ICCs calculated for measures repeated 28 days or more apart ([312], [313], [314], [307], [315], [316] and [317]). Six of these studies had lower ICCs than the overall pooled estimates, suggesting that variation of depression measurement instruments is higher than studies looking only at short-term test-retest reliability might suggest. The majority had time intervals between tests of one to two weeks, making them poor measures of long-term variation. For more accurate estimates of how the results of these tests vary over time, more repeat measures, over longer time scales, in more individuals is needed. It is also of note that the variation of the measurement instruments appears to be greater in patients with mental health diagnoses than with physical health diagnoses, although the number of included studies with mental health diagnoses is too small to draw firm conclusions.

Even for the main aims of these individual studies, there are some weaknesses. There are at least ten ways of calculating an ICC, all of which can give different results[153]. Only ten of the studies included in this review provide the method by which they calculate ICC, which limits the ability to interpret their results. None of the studies were blinded, (it was not ensured that the professionals administered the tests and assigned the scores without knowledge of previous

patient scores). This may not have had a marked effect on the test results since the questionnaires are standardised and there is no possibility of interpretation by the investigator regarding the numerical scores. However, it is possible that the subjects remembered their answers from their prior tests which could influence their scores [318], especially where the interval between tests was short. Information on ethnicity was often not recorded.

The patient populations of the included studies were diverse in terms of age, ethnicity, sex and health status. Many of the studies provided analysis of the variation of the outcome measures in subgroups. To choose which group to include in the meta-analysis was problematic, since it had to involve arbitrary choices. However, the rules to decide which group to include were applied systematically, using the following hierarchy: 1. Healthiest population; 2. Most stable population; 3. First listed population. There were missing values in a number of studies for age, sex and ethnicity, with ethnicity being particularly poorly recorded.

The limited number of measures and the short time periods between measurements mean that the current literature provides very limited information on how measures of depression symptoms vary over time within an individual. This lack of knowledge of how depression measures vary could make clinical decision-making difficult. If a test of depression symptoms with high within-individual variation is performed once by a clinician, it will be unclear whether the result obtained will represent a good estimate of the true mean state of that individual, or whether due to random changes in situational stressors or other temporary factors the measurement obtained is in reality markedly lower or higher than the true mean. This could lead to inappropriate influence on clinical management decisions, even when other clinical factors are taken into account. It is surprising that none of the studies identified

specifically related to patients with depression.

Studies are therefore needed that more accurately demonstrate the within-individual variation of tests of depression over time. Studies on depression measurement instruments such as those in this review with multiple measures over longer periods of time would demonstrate the total within-individual variation of these tests, encompassing the measurement error and the biological variation. This total variation is the most useful measure for clinicians since it provides guidance on the “real-world” variation of the selected tests, and hence the best idea of whether a single measurement can be relied upon, and if not, how many measurements are required to make meaningful clinical decisions. Studies assessing the variation of measurements of depression in patients with depression are currently lacking and are needed.

Note that some of the underlying diseases for which tests are performed have a higher variability than others, that is pathological variation will comprise a higher proportion of the total variability.

Limitations

There are a few limitations with this review. One is the possibility of relevant studies being missed by the search strategy. Since few of the studies were interested in measuring variation as a primary objective, but were focussed mainly on reliability of the tests, there may have been studies that report variation as a secondary outcome that have not been included. However, the search strategy was as broad as possible, and a large number of abstracts and studies were screened. On the other hand, this broad search strategy could have led to the inclusion of poor-

quality studies. There is no risk of bias tool currently in existence that is ideal for this review, but it was decided to use a validated risk of bias tool instead of a bespoke tool, and the COSMIN tool was selected. This enabled some estimate of risk of bias of each paper, although the questions on blinding were problematic, given that none of the studies included stated that the investigators performed the measurements or assigned scores without knowledge of previous scores by the individual.

Nevertheless, since the questionnaires were largely self-administered by the subjects, and provided numerical scores, this was unlikely to bias the findings significantly. The risk of bias scores are therefore presented in this study with the two questions on blinding excluded. Risk of bias tools are discussed further in Chapter 2.6.

5.7 Conclusion

This comprehensive systematic review shows that the published data for variation of depression symptom measurement instruments is very sparse and of poor quality and generalisability. Where variation data was published it was almost all in the form of test-retest reliability studies and reported in terms of ICC. Generally these studies showed excellent reliability. Studies in healthy subgroups showed lower ICC than those with physical illness.

I have shown in my systematic reviews that the previously published data on within-individual variation of HbA1c, CRP, respiratory function tests and depression symptom measurement scores are limited in terms of number of studies, quality of studies and generalisability. In the next four chapters, I therefore estimate the real-

world within-individual measured variation of these tests using large databases of routinely collected medical data in primary care databases. The following four studies will involve data from routine clinical practice, rather than a research context, and will utilise the actual measurements on which clinicians base their decisions. I will investigate two laboratory tests (HbA1c and CRP), the respiratory function tests FEV1 and FVC and the depression symptom measurement instruments PHQ-9, BDI and HADS, in retrospective cohort analyses. The first three of these studies involve data from the IMRD database. The next chapter shows the estimate of HbA1c variation.

6 Chapter six: Within-individual Variation of HbA1c

measurements in primary care: a retrospective cohort study

6.1 Abstract

Background

HbA1c is a marker for diabetes mellitus that reflects average glucose concentrations over the previous eight to twelve weeks. It is used to aid in the diagnosis and management of diabetes. Variation in within-individual measured HbA1c may affect its clinical utility but estimates of this are based on limited data that is often not generalisable to real-world settings.

Methods

Data on HbA1c results and sociodemographic, lifestyle and comorbidity covariates were extracted from the IQVIA Medical Research Database (IMRD) database using the DEXTER tool. A minimum of four measurements in the same individual was the only inclusion criterion. Within-individual measured variation was calculated as a coefficient of variation (CV) using a linear regression random effects model for the whole population and various subgroups.

Results

587,023 participants were included in this study, making it the largest study of variation of HbA1c to date. The overall measured within-individual coefficient of variation (CV_T) was 0.20 (95%CI 0.20 to 0.20). This is around three times higher than reported in a previous systematic review, which had an overall CV of 0.070. CV_T increased with patient mean HbA1c level.

Strengths and limitations

The large number of participants and the real-world nature of the results are important strengths of this study. Weaknesses included difficulties accounting for confounding by indication.

Conclusions

Estimated within-individual variation in this analysis of real-world data is very high and is higher than previously reported. Variation increases with patient mean HbA1c i.e. with more severe disease status. This has important implications for the diagnosis, monitoring and clinical decision-making for diabetes.

6.2 Introduction

Diabetes mellitus (DM) is a chronic metabolic disorder characterised by disturbances of carbohydrate, fat and protein metabolism [319]. Type 1 diabetes mellitus (T1DM) accounts for around 5 to 10% of cases, and results from autoimmune destruction of the beta cells of the pancreas, causing an absolute insulin deficiency. It tends to have a younger age of onset. Type 2 diabetes mellitus (T2DM) accounts for 90 to 95% of cases. It has a complex aetiology, and results in

insulin resistance and relative insulin deficiency. Prediabetes refers to individuals who have glucose levels higher than normal but do not meet the criteria for diabetes [320]. Five hundred and thirty-seven million adults were estimated to be living with diabetes in 2021, estimated to rise to 783 million by 2045. Diabetes was associated with 6.7 million deaths in 2021 [321]. The global prevalence of prediabetes was estimated to be 464 million in 2021, predicted to increase to 638 million in 2045 [322].

Haemoglobin A1c (HbA1c) is formed from the non-enzymatic attachment of glucose to haemoglobin, and reflects average glucose levels over the preceding eight to twelve weeks [323]. It was previously commonly measured using National Glycohemoglobin Standardization Program (NGSP) reported in units of %, but more recent guidelines recommend it is measured using International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) methods and reported in units of mmol/mol or in % derived from the established master equations [324].

HbA1c levels of 6.0% to 6.4% (42–47 mmol/mol) are consistent with a diagnosis of prediabetes, although the American Diabetes Association uses the lower value of 5.7% (39mmol/mol) [125, 325]. Levels of 6.5% (48mmol/mol) or more are consistent with a diagnosis of diabetes, although age, ethnicity, gestational status and genetic factors can affect HbA1c levels [124].

However, HbA1c levels, like all biological measurements, vary within an individual over time. This within-individual variation is broadly composed of three parts. Pre-analytical variation is due to variation of factors prior to a measurement, such as exercise, recent food or fluid intake or stress. Analytical variation is the variation introduced by imprecision in the measuring process.

Biological variation is the within-individual variation, influenced partly by predictable factors such as season, time of day or monthly hormone cycles, but also to a large extent by chance. The sum of these three types of variation gives the total within-individual variation [15]. The higher the within-individual variation, the lower the probability that a single measurement is reflective of the true mean. In this study, it is not possible to separate analytical from biological variation, so within-individual coefficient of variations (CV) is taken to refer to total within-individual variation i.e. the sum of biological and analytical variation.

The EFLM variation database lists the variation of common, laboratory-measured clinical tests [326]. However, most of its data comes from trials with small numbers of participants, often measured in ideal conditions. A recent systematic review ([327] and chapter two) found 111 studies reporting variation of HbA1c in a total of 564,143 participants. But most studies included few participants and/or small numbers of repeat measurements. This review also found that variation increases with mean HbA1c, and was higher in participants with T1DM and T2DM. However the extent to which HbA1c variation differs between different subgroups such as age, health status, BMI and ethnicity is not known.

NICE guidelines suggest diabetes can be diagnosed with a single HbA1c measurement in patients with appropriate symptoms, although they suggest a repeat test is “sensible.” In asymptomatic patients they recommend a second HbA1c measurement [328]. However, this advice only applies to initial diagnosis, not monitoring. Published studies on longitudinal testing of HbA1c tend to focus on an adequate number of tests being performed in a year, rather than repeat tests in the short-term to ascertain the accuracy of the test result, and in fact they sometimes

actively discourage short-term repeat testing as wasteful of health resources [329, 330].

This study aimed to use real-world data (that is data from a clinical source rather than an experimental setting) from the IQVIA Medical Research Database (IMRD) to estimate total within-individual variation of HbA1c in patients with and without diabetes and to compare this to previously published data. It explores the extent to which HbA1c variation differs with age, health status, BMI, ethnicity and other characteristics. It also considers the clinical implications of HbA1c variation in relation to diagnosis, monitoring and management of diabetes.

6.3 Aims

To describe the within-individual variation of HbA1c in a large cohort of patients in a database of routine general practice records, and to describe any covariates that have an effect on the variation. To describe the probability of within-individual variation in HbA1c resulting in clinically important changes in HbA1c.

6.4 Methods

6.4.1 Study design and data sources

A population-based retrospective cohort study was undertaken using data obtained from the IQVIA Medical Research Database (IMRD). IMRD incorporates data from The Health Information Network (THIN), a Cegedim Database. Reference made to THIN is intended to be descriptive of the data asset licensed by IQVIA.

IMRD is a pseudonymised database of the electronic healthcare records of patients registered with general practices in the UK which use compatible practice management systems. As of March 2021, the database contained the record of over 20 million patients, of which around 4 million were active with their registered GP practice (<https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/the-health-improvement-network-thin-database/>).

IMRD data has been shown to be generalisable to the general primary care population in terms of factors such as demographics, deprivation, condition prevalence and deaths [117]. The dataset used for this study consisted of 11,737,653 registered in 832 practices.

This study is reported according to the STrengthening the Reporting of Observational studies in Epidemiology (STROBE) statement for observational studies.

6.4.2 Ethical approval

IQVIA Medical Research data has been approved by the NHS Health Research Authority (NHS Research Ethics Committee ref 18/LO/0441) for the purpose of medical and public health research and to supply data for researchers for scientifically approved studies. The use of the data for this study was approved by the IQVIA Scientific Review Committee on 19th November 2020 with the reference 20SRC068.

6.4.3 Setting

Data was obtained from UK general practices registered with THIN/IMRD from database inception to 9th September 2021. Data extraction from the database was performed using the Dexter software [122] and was completed on 9th September 2021. The date of the latest entrance was 11th January 2021.

6.4.4 Participants

All participants in the database were eligible for inclusion. Except for sensitivity analyses, participants were excluded if they did not have at least four repeat measurements of HbA1c expressed in mmol/mol recorded. Four was chosen as the cut-off for repeat measurements for pragmatic reasons related to maximising the number of repeat measurements in the analysis while allowing a sufficiently large sample size for analyses. Note this was a higher minimum number of repeat measures than for the systematic reviews, since the datasets used in this study were much larger, allowing participants with less than four repeat measurements to be excluded while still being able to include large numbers of participants in the analyses. Participants were followed until the earliest date of either death, leaving the practice, stopping contributing to the database or the study end date.

6.4.5 Variable choice

The primary outcome was within-individual variation of HbA1c calculated as the total within-individual coefficient of variation (CV_T). Data on a number of other patient and test characteristics were also extracted: socio-demographic characteristics,

diabetic status, diagnoses and comorbidities. Variables extracted that are included in the analyses are listed in table 6.1 Variables were selected as those I hypothesised most likely to have an effect on variation or that enabled calculation of variation (e.g. date of measurement). Code lists for the different variables are available at: <https://github.com/THINKINGGroup/phenotypes>.

Table 6.1 Variables extracted from database.

Variables extracted	
Test characteristics	HbA1c result
	Date of measurement
	Unit of measurement
Patient characteristics	
Sociodemographic	Age
	Sex
	Ethnicity
	Townsend Deprivation Score
	BMI
	Geographical region
Lifestyle factors	Smoker status
	Alcohol consumption status
Diabetic status	Prediabetes
	Type one diabetes mellitus
	Type two diabetes mellitus
	Diabetic medications
Diagnoses and comorbidities	Hypertension
	Hyperthyroidism
	Hypothyroidism
	Ischaemic heart disease
	Heart failure
	Ischaemic stroke
	Haemorrhagic stroke
	All cancers

6.4.6 Bias

The study size included the entire database where participants had sufficient repeat measurements to reduce bias and to ensure that the sample sizes for the

whole population, subgroup analysis and sensitivity analyses were sufficiently large to compensate for any weaknesses inherent in routinely collected medical data.

6.4.7 Study and subgroup sample size

All participants with sufficient repeat HbA1c measurements of HbA1c were included. The minimum number of participants for subgroup analysis was selected as 100, for pragmatic reasons, compromising between group size and number of subgroup analyses possible. Using simulation, assuming a CV of 20%, with 100 participants and 4 observations per participant provides an estimate with an approximate 95% confidence interval with a total width of less than 10% (+/- 5%).

6.4.8 Data cleaning

When calculating coefficient of variation, mmol/mol and percentage are not interchangeable since one measure has an intercept on the y axis when converting to the other. Since most HbA1c measurements were reported in mmol/mol, those expressed as a percentage were excluded from the main analysis. Values of HbA1c greater than 195.1 mmol/mol were excluded as being implausible based on analysis of the IMRD data where fewer than 0.01% of measurements had values greater than 195.1mmol/mol. Values of less than 20.1 mmol/mol were excluded as either implausible based on published data, or because they were possibly due to some unknown biological mechanism unrelated to diabetes [331]. However, a histogram of values <20.1 mmol/mol had a normal distribution around a median of 7.7mmol/mol,

values more consistent with units of per cent, suggesting most of these low values were actually percent values erroneously recorded as mmol/mol. For BMI, implausible values were considered <14 or >70 [332] and these were re-categorised as missing.

6.4.9 Statistical methods

Statistical calculations were performed using Stata 17.0. Within-individual variance was estimated using a random effects model ('mixed' command). A coefficient of variation was then calculated using the formula $CV = \text{standard deviation} / \text{mean}$, expressed as a percentage or a decimal fraction. Since HbA1c values are often expressed as a percentage, it was elected to express the CV as a decimal fraction in this study. 95% confidence intervals were estimated from the model in Stata. Participants with identical HbA1c results on at least four different occasions were considered improbable and likely duplicated results, so were excluded. If more than one measurement was performed on the same day, only the first result was included on the basis that others recorded for the same day were likely to be duplications or errors.

Minimal clinically important difference (MCID) was considered to be 5.5 mmol/mol (0.5%) [333]. The chance of getting a difference of at least this amount between the first and second results was calculated as a percentage. This was achieved by counting the number the number of participants in which every combination of first and second results was achieved. The probability of each combination of first and second results was then calculated by dividing the number of participants with that combination by the total number of participants who had the first result.

Subgroup analyses were performed by repeating the analysis on subsets of the data: by age, sex, ethnicity, Townsend neighbourhood deprivation score quintiles, (which give a score from 1 to 5 estimating deprivation in a local area, with 1 being the least deprived and 5 being the most deprived [334]), smoking and alcohol status, geographical location, year, quarter, time between measurements, mean HbA1c result, presence of prediabetes, diabetes mellitus types 1 and 2 and other comorbidities and use of diabetic medication. Further subgroup analyses were performed for participants divided according to the value of multiple variables (e.g. sex male, ethnicity white, age young, no diabetes), with the variables selected being those that had the most important effects on variability based on univariable analysis and which I hypothesised to be most likely clinically to have an effect.

Age and sex data were complete. Where a code for a comorbidity was absent, it was assumed not to be present. For missing ethnicity or deprivation score a missing category was present in the data. For BMI, the data appeared to be complete but in fact had a large number of implausible values – these were transformed into a missing category (see above). The analysis of subgroups based on multiple variables included only those without missing data.

Sensitivity analyses were performed to assess bias due to factors which might affect the robustness of the results such as which statistical methods were used, number of repeat measurements included and unit of measurement. Therefore analyses by year and quarter were performed to look for changes in population variation over time or with season. Analyses were also performed by unit (mmol/mol v percentage) to assess for differences in variation due to method or mathematical factors, by number of measurements for each individual patient (calculating the CV for patients with 2, 3, 4 etc measurements each) and by inclusion or exclusion of

implausible values. A simple arithmetic method of calculating CV was also performed in which the standard deviation of each individual and the mean of each individual were used to calculate an individual CV_T , which was then averaged using the mean. CV was also calculated using log-transformed data.

Statistical code for Stata is available from the author on request.

6.5 Results

Participants

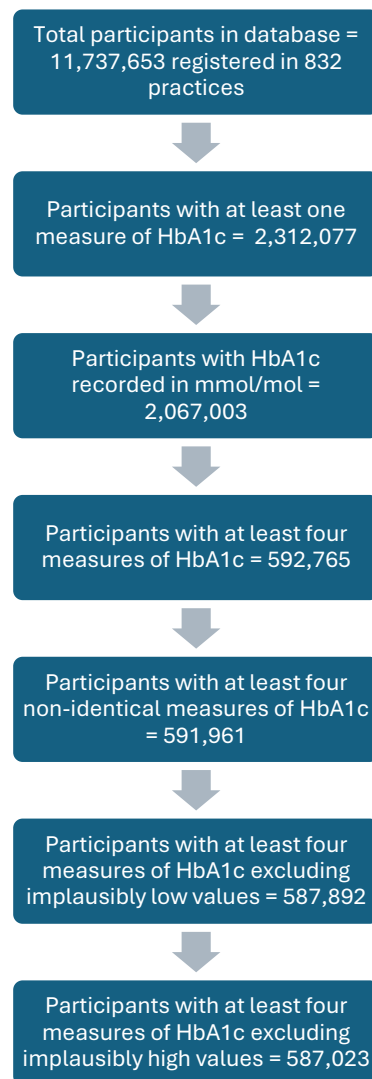


Figure 6.1 Participant selection algorithm

11,737,653 participants registered in 832 practices were available in the database for analysis. Figure 6.1 shows the flowchart of participant selection. 587,023 participants had at least four measures of HbA1c recorded as mmol/mol with values between 20.1 (4%) and 195.1 (20%) mmol/mol. The median number of measurements per individual was 11 (IQR 7-18).

6.5.1 Descriptive data and outcome data

Table 6.2 shows the participant characteristics and main univariable subgroup analyses. 587,023 participants were included in the study, with a median age at first measurement of 70 (IQR 58 to 79), 51.7% being male. The median number of years between first and last measurement in the database was 5 (IQR 3 to 7).

The overall coefficient of variation for HbA1c was 0.200 (95% CI 0.200 to 0.200), with a mean HbA1c of 52.31 mmol/mol. Variation increased with mean HbA1c levels up to moderately high mean HbA1c levels and with median HbA1c to markedly high median HbA1c levels (figs 6.3 and 16.13).

Tables 6.2 and 16.1 shows the variation of HbA1c according to diabetic status and diabetic medication. Variation was much higher in participants with prediabetes or diabetes compared to those without, and much higher in participants treated with anti-hyperglycaemic drugs than those not treated with these drugs.

Diabetic status, deprivation score, ethnicity, sex and age appear to have important effects on CV, so these were chosen for multi-subgroup analysis. The results of this are presented in a heat map in fig 6.2. It can be seen from this that of these subgroups, diabetic status seems to have the most marked effect on CV_T.

Table 6.2 Sociodemographic data and main univariable subgroup analyses.

		N (%)	Mean	CV _T (95% CI)
All		587023 (100)	52.31	0.200 (0.200 to 0.200)
Sex	Male	303,622 (51.7)	53.71	0.201 (0.201 to 0.202)
	Female	283,401 (48.3)	50.80	0.198 (0.198 to 0.199)
Age	<=10	259 (0.0)	62.23	0.178 (0.171 to 0.184)
	>10 to 20	1,949 (0.3)	63.74	0.209 (0.205 to 0.212)
	>20 to 30	7,474 (1.3)	57.77	0.218 (0.125 to 0.220)
	>30 to 40	18,434 (3.1)	52.66	0.225 (0.224 to 0.227)
	>40 to 50	45,032 (7.7)	52.75	0.225 (0.224 to 0.226)
	>50 to 60	97,285 (16.6)	53.95	0.221 (0.220 to 0.221)
	>60 to 70	133,560 (22.8)	52.75	0.206 (0.205 to 0.206)
	>70 to 80	152,239 (25.9)	51.11	0.188 (0.188 to 0.189)
	>80 to 90	101,672 (17.3)	51.26	0.180 (0.180 to 0.181)
	90 to 100	27,945 (4.8)	51.79	0.179 (0.178 to 0.180)
Ethnicity	Black	10,971 (1.9)	53.16	0.228 (0.226 to 0.230)
	Mixed	2,303 (0.4)	51.91	0.212 (0.208 to 0.216)
	Others	8,221 (1.4)	52.66	0.192 (0.190 to 0.194)
	South Asian	25,195 (4.3)	53.07	0.184 (0.183 to 0.184)
	White	304,062 (51.8)	51.97	0.201 (0.201 to 0.201)
	Missing	236,271 (40.2)	52.62	0.200 (0.200 to 0.200)
BMI	<18.5	3,531 (0.6)	47.28	0.202 (0.199 to 0.205)
	18.5 to <25	107,222 (18.3)	48.63	0.181 (0.181 to 0.182)
	25 to <30	179,744 (30.6)	51.58	0.189 (0.189 to 0.189)
	30 to <35	116,600 (19.9)	54.48	0.201 (0.201 to 0.201)
	>35	80,052 (13.6)	56.98	0.217 (0.216 to 0.217)
	Missing	99,874 (17.0)	51.40	0.207 (0.207 to 0.208)
Deprivation score	1	103,029 (17.6)	51.20	0.185 (0.185 to 0.185)
	2	99,045 (16.9)	52.14	0.191 (0.190 to 0.191)
	3	104,133 (17.7)	52.79	0.199 (0.198 to 0.199)
	4	96,380 (16.4)	53.83	0.207 (0.206 to 0.207)
	5	72,831 (12.4)	54.66	0.215 (0.215 to 0.216)
	Missing	111,605 (19.0)	50.19	0.204 (0.203 to 0.204)
Diabetic Status	No DM	148,870 (25.4)	38.24	0.091 (0.091 to 0.091)
	PreDM	73,649 (12.5)	48.74	0.180 (0.179 to 0.180)
	T1DM	33,404 (5.7)	70.17	0.157 (0.156 to 0.157)
	T2DM	329,038 (56.1)	58.70	0.194 (0.193 to 0.194)

6.5.1.1 Subgroup analyses

Higher CV_Ts were seen in the following subgroups: males compared to females; ex-drinkers compared to current drinkers or teetotallers; current smokers compared to ex-smokers or participants who have never smoked; middle-aged compared to older or younger participants; obese and underweight participants compared to normal weight and overweight participants, participants of black ethnicity compared to other ethnicities. Little difference in CV_T was seen between patients with different comorbidities except prediabetes or diabetes.

Tables 6.2 and Appendix 6.1 table 16.1 and figs 16.1 to 16.9 show the major subgroup analyses.

Diabetic status	Deprivation	Sex Age Ethnicity	Male				Female			
			20-60		>60		20-60		>60	
			White	Non-white	White	Non-white	White	Non-white	White	Non-white
BMI										
No diabetes	1 to 3	<30	9.5	12.3	7.7	7.7	6.2	7.8	6.5	7.2
		>30	13.1	13.4	11.1	9.5	9.5	11.2	8.1	8.8
	4 to 5	<30	10.9	12.9	8.1	9.7	9.8	8.5	8.3	7.9
		>30	17.3	14.2	15.3	12.3	11.0	10.6	9.8	9.6
Diabetes	1 to 3	<30	17.6	12.8	11.5	12.1	15.8	14.0	9.0	10.7
		>30	19.5	20.6	14.3	13.9	16.4	17.9	12.9	10.1
	4 to 5	<30	18.4	17.5	10.7	11.6	15.5	13.6	10.3	11.0
		>30	23.0	19.2	18.3	16.6	19.3	18.1	11.4	14.6

Figure 6.2 Heat map showing the coefficient of variation expressed as % for multiple subgroup analysis according to sex, age, ethnicity, deprivation and BMI.

Age is in years; deprivation score is Townsend deprivation score.

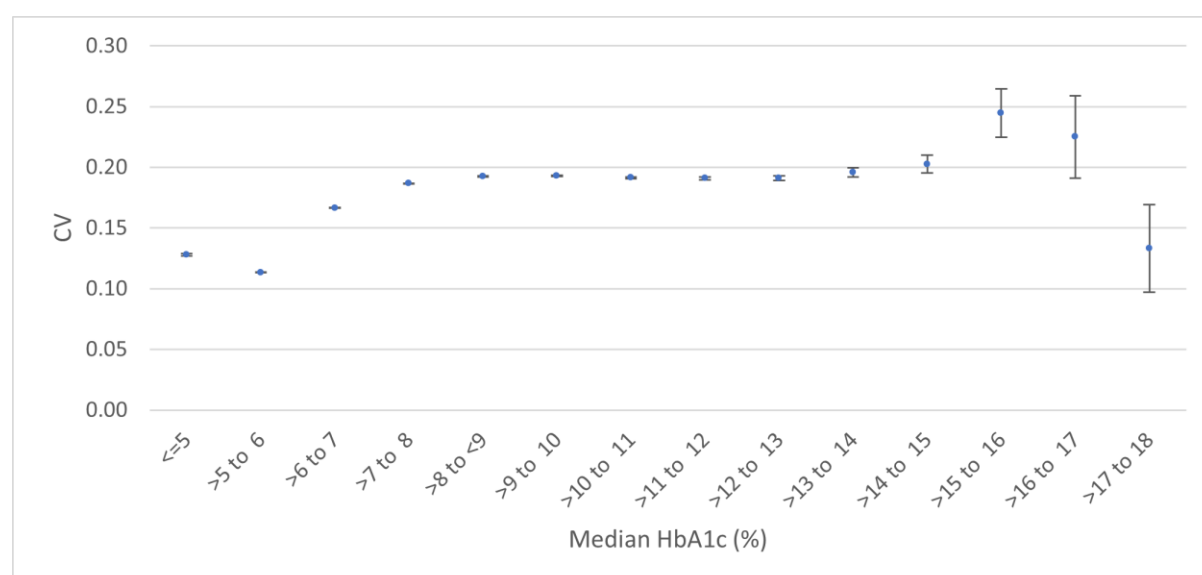


Figure 6.3 Coefficient of variation by patient median with 95% confidence intervals

first result	26 to 31	>31 to 37	>37 to 42	>42 to 48	>48 to 53	>53 to <58	>58 to 64	>64 to 69	>69 to 75	>75 to 80	>80 to 86	>86 to 91	>91 to 97	>97 to 102	>102 to 108	>108 to 113	>113 to 119	>119 to 124	>124 to 130	>130 to 135	>135 to 140	%	Difference -0.5% or less	Difference +5% or more
26 to 31	47.85	40.45	7.47	0.87	0.35	0.15	0.09	0.22	0.07	0.05	0.01	0.04	0.04	0.04	0.01	0.04	0.01	0.00	0.00	0.00	0.00	4.5 to 5		49.93
>31 to 37	7.53	57.88	31.43	1.82	0.44	0.26	0.08	0.09	0.05	0.06	0.05	0.02	0.02	0.03	0.02	0.03	0.01	0.01	0.00	0.01	0.01	>5 to 5.5	7.53	34.42
>37 to 42	0.52	12.48	65.29	16.52	2.96	0.94	0.31	0.25	0.12	0.11	0.07	0.06	0.06	0.07	0.05	0.05	0.02	0.02	0.01	0.01	0.01	>5.5 to 6	13.00	21.66
>42 to 48	0.08	0.94	23.74	47.49	19.32	4.68	1.22	0.85	0.33	0.32	0.17	0.12	0.12	0.14	0.09	0.10	0.05	0.04	0.03	0.02	0.02	>6 to 6.5	24.77	27.63
>48 to 53	0.06	0.30	4.34	21.24	43.77	20.12	4.42	2.38	0.91	0.76	0.40	0.21	0.21	0.24	0.13	0.12	0.06	0.06	0.04	0.03	0.02	>6.5 to 7	25.94	30.13
>53 to <58	0.05	0.23	1.92	6.59	22.24	39.24	14.52	7.90	2.60	1.85	0.76	0.38	0.38	0.32	0.17	0.17	0.09	0.08	0.06	0.04	0.03	>7 to 7.5	31.03	29.35
>58 to 64	0.04	0.22	1.64	4.06	10.59	26.11	24.74	17.51	6.12	3.98	1.54	0.60	0.60	0.51	0.29	0.27	0.14	0.13	0.06	0.04	0.04	>7.5 to 8	42.67	31.82
>64 to 69	0.03	0.21	1.34	3.11	7.07	16.31	17.52	26.49	11.25	7.99	3.00	1.05	1.05	0.85	0.43	0.37	0.18	0.14	0.09	0.09	0.05	>8 to 8.5	45.59	26.54
>69 to 75	0.08	0.22	1.36	2.56	5.46	11.05	12.04	19.86	17.67	14.71	5.64	1.75	1.75	1.32	0.53	0.52	0.32	0.26	0.08	0.07	0.03	>8.5 to 9	52.62	26.97
>75 to 80	0.02	0.26	1.24	2.43	4.62	8.27	8.81	15.77	13.75	20.18	8.93	3.06	3.06	2.33	1.01	0.69	0.45	0.33	0.15	0.11	0.09	>9 to 9.5	55.18	20.22
>80 to 86	0.07	0.34	1.59	2.67	4.58	6.74	7.34	10.75	9.85	16.68	14.90	4.58	4.58	3.68	1.85	1.52	0.49	0.41	0.28	0.13	0.14	>9.5 to 10	60.63	17.66
>86 to 91	0.04	0.26	1.61	2.68	4.42	6.49	5.64	9.49	7.87	13.19	10.71	7.72	7.72	5.96	2.56	1.88	0.76	0.70	0.28	0.23	0.14	>10 to 10.5	62.39	20.23
>91 to 97	0.08	0.26	1.99	2.76	4.23	6.55	5.50	7.99	6.98	10.08	9.01	12.52	12.52	8.83	3.99	2.90	1.22	0.92	0.37	0.29	0.20	>10.5 to 11	67.93	18.71
>97 to 102	0.05	0.33	1.58	2.48	4.12	6.24	5.72	7.65	6.53	8.83	7.35	9.24	9.24	13.72	5.51	4.42	1.91	1.45	0.51	0.49	0.15	>11 to 11.5	69.36	14.44
>102 to 108	0.07	0.22	1.85	2.85	3.87	6.92	5.00	6.52	5.38	8.12	6.67	7.43	7.43	9.92	11.01	6.63	3.20	2.34	0.87	0.56	0.49	>11.5 to 12	72.24	14.08
>108 to 113	0.09	0.30	2.26	3.38	5.02	6.13	5.00	6.86	4.98	7.16	5.30	6.05	6.05	9.15	7.18	11.92	4.10	3.72	1.32	1.07	0.75	>12 to 12.5	74.89	10.96
>113 to 119	0.08	0.45	2.37	3.12	5.26	6.88	4.85	5.94	4.96	6.39	5.30	5.41	5.41	7.82	6.05	10.90	8.76	4.74	2.18	1.01	0.49	>12.5 to 13	81.17	8.42
>119 to 124	0.04	0.36	2.20	3.88	4.68	6.04	5.20	6.77	4.92	5.04	4.80	5.08	5.08	7.21	4.48	8.37	5.40	11.73	3.60	1.76	1.28	>13 to 13.5	79.58	6.65
>124 to 130	0.07	0.37	2.79	3.30	5.36	6.83	5.80	6.39	5.14	6.24	3.96	4.55	4.55	5.36	4.70	6.09	4.77	6.61	7.86	4.33	1.84	>13.5 to 14	82.89	6.17
>130 to 135	0.00	0.10	2.97	2.97	4.21	6.60	4.21	5.84	5.17	6.99	3.54	4.31	4.31	5.45	3.54	6.79	3.92	6.60	5.45	8.80	3.44	>14 to 14.5	82.97	3.44
>135 to 140	0.00	0.12	1.96	2.94	4.53	6.00	5.64	6.25	4.29	6.00	3.92	4.78	4.78	4.53	3.55	6.13	4.41	5.39	3.31	6.13	7.48	>14.5 to 15	84.68	
%	4.5 to 5	>5 to 5.5	>5.5 to 6	>6 to 6.5	>6.5 to 7	>7 to 7.5	>7.5 to 8	>8 to 8.5	>8.5 to 9	>9 to 9.5	>9.5 to 10	>10 to 10.5	>10.5 to 11	>11 to 11.5	>11.5 to 12	>12 to 12.5	>12.5 to 13	>13 to 13.5	>13.5 to 14	>14 to 14.5	>14.5 to 15			

Figure 6.4 Probability of second result being 5 mol/mol or more different than first result.

The last two columns show the probability of a second result being at least 0.5% less or more than the previous result.

Fig 6.4 shows a heat map with data on the probability of the result of a second HbA1c test given the result of an initial HbA1c test.

6.5.1.2 *Sensitivity analyses*

Sensitivity analyses were performed for CV_T by year, quarter, number of measurements, using different thresholds for removing participants with implausible values, median HbA1c level, log mean HbA1c level, number of days between measurements, number of days between measurements stratified by diabetes status, results coded as mmol/mol v % and a crude estimate of CV_T using a simple calculation for each individual rather than a linear regression model for the group (See Appendix 6.2 table 16.2 and figs 16.10 to 16.19). There was no evidence of marked seasonality or a trend over years for the CV_T . An increase in CV_T with number of measurements was seen up until approximately 9 repeat measurements after which the CV_T plateaued. No major differences were seen using different values of cut-off points for implausible values. There was no major difference between CV_T stratified by mean or median, except at extremely high or low values where median CV_T was higher than mean, and no major difference when the data was log-transformed. CV_T decreased with increasing median days between measurements, but when this was stratified by diabetic status, CV_T decreased with increasing median days between measurements for non-diabetic participants but not for diabetic participants. The CV_T was higher for participants with HbA1c reported as mmol/mol than as %, but if the mmol/mol were converted into % and then the CV_T was calculated, the CV_T between this group and the group originally reported as % was similar. This is probably because coefficient of variation is not invariant to transformations that involve adding or subtracting a constant [335]. Calculation of

CV_T by a crude arithmetic method resulted in a lower value than using the linear regression method.

6.6 Discussion

6.6.1 Key results

This is the largest single study of variation of HbA1c, with 587,023 participants, more than all the participants put together in the recent systematic review of 111 studies ([327] and chapter two). The overall total within-individual coefficient of variation (CV_T) for this population was 0.200 (95% CI 0.200 to 0.200) which is almost three times higher than the median reported in the systematic review (0.070). This study found the CV_T for non-diabetic patients was 0.091 (0.091 to 0.091), over five times higher than in the systematic review (0.017) and the EFLM biological variation database (0.016) [326]. In this study, participants with diabetes or prediabetes had much higher variation than those without these conditions. For patients with T1DM, the CV_T of HbA1c in this study was 0.157 compared to 0.084 for the systematic review, and for T2DM the CV_T of HbA1c was 0.194 compared to 0.083 in the systematic review - for both types of diabetes a finding of roughly twice the variation of HbA1c compared to that previously published. Note that the overall CV_T includes within-individual variation in patients before and after diagnosis of DM, so it includes the long-term increase in HbA1c from non-diabetic to diabetic status. It is therefore higher than when within-individual variation is calculated for a single category of diabetic status. Total within-individual variation (CV_T) by diabetic status may be a more useful measure than biological variation (CV_I) when considering clinical implications.

CV_T increased with the median HbA1c (fig 6.3) up to values of around 75mmol/mol (9%). This is important because patients with more severe disease or who are less well-controlled may have a higher variation, making management recommendations more challenging in this subgroup. Participants treated with anti-hyperglycaemic drugs had much higher variation than those not treated with these drugs. Male sex, black ethnicity, ex-drinkers, current smokers, high BMI and higher deprivation levels were associated with higher HbA1c variation. The increased variation in ex-drinkers may be caused by those who had to give up drinking for health reasons. Smoking is known to be associated with risk of diabetes onset and complications [336] Sex, BMI and deprivation may be independent risk factors for increased variability and there is evidence that ethnicity and social factors affect diabetes outcome [337] but there may be some interaction between the subgroups, for example BMI, deprivation and diabetic status being interdependent. It should also be noted that the median age of subjects was quite high at 70, suggesting this population may be more prone to pathological variation due to an increased risk of diabetes and other diseases. Multiple subgroup analysis (figure 6.2) shows that diabetic status is the most important determinant of variation. Figure 6.4 shows the probability of a second measurement being different from the first measurement by 5 mmol/mol (approximately 0.5%) – the minimum clinically important difference (MCID). This shows that for every plausible initial value of HbA1c it is common for the second measurement to differ from the first by at least the MCID.

The CV was higher when HbA1c was reported as mmol/mol compared to reported as %, but if CV was calculated after converting mmol/mol to %, the CV_T between the two % groups was similar. The difference between CV reported in % compared to mmol/mol is probably related to the fact that the conversion equation

has an intercept, since the intercept does not vary. As noted in the results section, the coefficient of variation is sensitive to transformations that involve adding or subtracting a constant [335]. This will be more important at low values – at higher values the effect of the intercept on the variation will be minimised. This problem with calculating the coefficient of variation on HbA1c reported in units of % does not appear to be widely recognised.

6.6.2 Strengths and limitations

The major strengths of this study are the large size of the dataset and that the data is real-world data taken from the clinical records of primary care practices in the UK, and not measurements taken under ideal conditions. The Biological Variation Data Critical Appraisal Checklist (BIVAC) recommends biological variation studies are performed on patients in steady state, with “preanalytical procedures described and standardized to minimize preanalytical variation” [154]. This is important when using biological variation data to set analytical performance specifications, but is less relevant to clinical decision-making. The real-world nature of the data is important because these are the results on which clinical decisions are made. As these data are the results seen by clinicians, they include every component of variation (pre-analytic, analytic, biological and pathological). It therefore includes variation due to changes in patient status such as diabetic status, weight and medication. This is less of an issue for those not on medication (e.g. non-diabetic or prediabetic) and those where measurements are closer together in time. The CV is actually higher in participants with measurements closer together in this study, but this may be an example of confounding by indication, i.e. participants that have less well-controlled diabetes are more likely to have more frequent measurements. Similarly, the higher variation seen in participants with higher mean HbA1c values or with diabetes could

also be linked to more frequent measurement. Nevertheless, the large disparity between this study and previously published results and the large number of participants helps to compensate for this problem, and it is biologically plausible that patients with diabetes and higher mean HbA1cs have higher variation.

Assumptions were made for the repeat results calculation, namely that the change between first and second result is representative of changes in the result in subsequent calculations, which may not be the case.

Another weakness with the data is error associated with data entry and recording into the clinical databases. Data cleaning prior to analysis partially compensates for this, (as well as the fact that most lab results are transmitted electronically). For example implausible values were removed, taking the cut-off points from published data (e.g. [331]) and analysis of the IMRD data itself, removing of duplicate results on the same day, elimination of participants with less than four repeated measurements, and then removing any patients with four identical results which is a highly unlikely finding. Note that removing implausible values reduces the estimated variation and yet the estimated variation remains very large.

Inclusion of the entire dataset meant the only potential bias introduced into the results (that was not already present in the database) could come from data cleaning (e.g. exclusion of implausible values) and the specification that at least four measurements must be included. However, this latter specification, although aimed at increasing the accuracy of the variation calculation, could introduce confounding by indication since patients with more frequent tests might have more severe/less stable disease and hence a higher degree of pathological variation. Furthermore, it has been shown that the ordering of any test by a clinician increases the risk of the

disease which the test is investigating, since the presence of a test in a patient record suggests there was a suspicion that patient was at risk of the disease. This could cause some degree of selection bias in the full dataset when compared to results obtained from healthy populations in controlled biological variation studies [338].

6.6.3 Interpretation and generalisability

This study shows that the estimated measured total within-individual variation of HbA1c in this population of primary care patients in the UK is much higher than results for variation previously reported in the literature. This may be because the participants were less likely to be in a steady state, and the pre-analytical testing conditions were not rigorously controlled. This means that the results are likely to be more generalisable to a primary care population.

The findings of this study have clinical implications for diagnosis and monitoring. NICE guidelines recommend a diagnosis of diabetes is made if appropriate symptoms and a single measurement of HbA1c of more than 48mmol/mol (6.5%) is present, or if asymptomatic, two measurements of HbA1c of more than 48mmol/mol (6.5%) [328]. However, as figure 6.4 shows, the chance of a patient with an initial HbA1c measurement of 48mmol/mol to 53mmol/mol (6.5% to 7%) having a second HbA1c level of 5.5mmol/mol (0.5%) less is 0.26, suggesting around one in four patients with an initial HbA1c consistent with diabetes will by chance have a second measurement not consistent with this diagnosis. False positives due to not taking within-individual variability into account could lead to a patient being incorrectly diagnosed with prediabetes (which might cause stress to the

patient and affect non-clinical aspects of the patient's life such as the ability to get insurance) or diabetes, which might lead to incorrect treatment. False negatives could lead to a delay in diagnosis and interventions which might increase the risk of progression and complications of the disease.

Although the management of diabetes is complex and requires multiple factors to be taken into account, in general guidelines recommend aiming for an HbA1c of less than 53mmol/mol (7%). The CV_T of this study can be used to calculate the probability with which a measured value reflects the true mean of the patient, (assuming normal distribution and using a Z-score of difference between means/standard deviation). If we take a hypothetical patient with a true HbA1c of 58mmol/mol (7.5%), the CV_T from the data in this study is approximately 0.182. This suggests a probability of 0.60 that the measured result will be 5.5mmol/mol higher or lower than the true mean, which is the minimal clinically important difference [333]. This could be the difference between recommending only lifestyle modifications and intensifying drug treatment according to NICE guidelines [339].

It has been previously shown that practitioners, particularly nurses, but to some extent doctors, are unaware of the inherent variation in HbA1c and adjust therapy based on very small changes in HbA1c [333]. It can thus be seen that not taking within-individual variation into account may lead to false conclusions about the magnitude and direction of change of HbA1c which could lead to appropriate interventions not being given if the measured value is higher than the real value by a clinically significant amount, leading to poor diabetes control. This would be particularly important if the real change was positive and the measured change was negative. Conversely, if the measured value was lower than the real value by a significant amount this could lead to interventions being given inappropriately with an

increased risk of adverse effects, and this is particularly important if the real change was negative and the measured change was positive.

This study emphasises that clinicians should be aware of the variation of HbA1c and it demonstrates a relatively high probability of misdiagnoses or incorrect inferences, for example with regard to management of diabetes when relying on single or small numbers of measurements of HbA1c. Further, authors of clinical guidelines should take variation into account when making diagnosis, monitoring and treatment recommendations, and researchers should be aware of the variation of HbA1c when assessing the magnitude of an effect in drug trials.

Note that some of the underlying conditions for which the tests examined in this thesis are utilised are much more variable than diabetes [340] (for example bacterial pneumonia and other acute illnesses, assessed using CRP testing). The pathological component of variation in different diseases therefore has a greater or lesser magnitude depending on the underlying disease, with diabetes being one of the conditions that has a comparatively lower degree of pathological variation.

Further research that addresses confounding by a change in frequency or timing of testing with increased disease severity or variation would be useful. Selecting a cohort with no change in medication may help evaluate variation in which there is no material change in the status of the participants. Studies into how variation in this population affects clinical outcomes such as mortality and whether reducing variation with improved testing and management protocols improves outcomes is also important. Modelling the interaction between different factors affecting variation such

as age, sex, BMI or diabetic status would also be helpful, but was beyond the scope of this paper.

6.7 Conclusions

The findings of this study suggest that the estimated measured total within-individual variation of HbA1c in the UK is much higher than results for within-individual variation previously reported in the literature. Within-individual measured total variation increased with median HbA1c up to values of around 75mmol/mol (9%).

Having estimated the within-individual measured variation of HbA1c and various subgroups, in the next chapter I perform the same analysis on CRP, since the systematic review has shown this to be a measurand with a high within-individual variability.

7 Chapter seven: Within-individual Variation of CRP

measurements in primary care: a retrospective cohort study

7.1 Abstract

Background

C-reactive protein (CRP) is a biomarker of inflammation used in diagnosis of inflammatory diseases and to guide treatment decisions. Variation in within-individual measured CRP may affect its clinical utility but estimates of within-individual variation are based on limited data and so may not be accurate.

Methods

Data on CRP results and sociodemographic, lifestyle and comorbidity covariates were extracted from the IMRD database using the DEXTER tool. A minimum of four measurements for each individual was the only inclusion criterion. CRP data were log-transformed for analysis. Within-individual measured variation was calculated as a coefficient of variation (CV) using a linear regression random effects model for the whole population and various subgroups.

Results

472,593 participants were included in this study, making it the largest study of variation of CRP to date by a factor of approximately five. The overall coefficient of variation for CRP was 1.148 (95% CI 1.147 to 1.149). This is much higher than the

median reported CV for CRP of previous studies which was 0.41. CV increased with patient median.

Strengths and limitations

The large number of participants and the real-world nature of the results are important strengths of this study. Weaknesses included difficulties accounting for confounding by indication, and the short half-life of CRP making it hard to distinguish between acute illness and physiological variation.

Conclusions

Estimated within-individual variation in this analysis of real-world data is very high and is higher than previously reported. Variation increases with patient median CRP, i.e. with more severe disease status. This has important implications for the diagnosis, monitoring and clinical decision-making for inflammatory disease.

7.2 Introduction

C-reactive protein (CRP) is an acute-phase protein synthesised in the liver, the plasma levels of which rise with inflammation [160]. It is a non-specific biomarker that aids with the diagnosis and assessment of severity of infectious and inflammatory diseases such as respiratory infections (including COVID-19) [341] and chronic inflammatory diseases such as rheumatoid arthritis [342]. It can help distinguish between aetiologies of disease, such as bacterial or viral pneumonia [343, 344]. Low levels of CRP can be assessed with the high-sensitivity CRP (hsCRP) test which has been shown to be useful as marker for cardiovascular disease [345].

Current National Institute for Health and Care Excellence (NICE) guidelines state that if it is unclear after initial assessment whether a patient with a lower respiratory tract infection needs antibiotics, CRP testing can be used to help with decision-making. It is recommended that after a point of care CRP test, patients with CRP more than 100mg/litre are given immediate antibiotics, patients with CRP less than 20mg/L are not routinely provided with antibiotics, and consideration is given to providing patients with CRP 20-100 mg/L with backup antibiotics. The American Gastroenterology Association guidelines for the role of biomarkers on the management of Crohn's disease suggest a cut-off of CRP of 5mg/L, below which active inflammation is unlikely and in certain circumstances advise clinicians use this cut-off to decide whether endoscopic evaluation should be performed [346].

However, CRP levels, like all biological measurements, vary within an individual over time. This within-individual variation is broadly composed of three parts. Pre-analytical variation is variation due to varying factors prior to a measurement, such as exercise, recent food or fluid intake or stress. Analytical variation is the variation introduced by imprecision in the measuring process. Biological variation is the within-individual variation, influenced partly by predictable factors such as season, time of day or monthly hormone cycles, but also to a large extent by chance. The sum of these three types of variation gives the total within-individual variation [15]. The higher the within-individual variation, the lower the probability that a single measurement is reflective of the true mean. In this study, it is not possible to separate analytical from biological variation, so within-individual coefficient of variations (CV_T) is taken to refer to total within-individual variation i.e. the sum of biological and analytical variation.

A database of biological variation exists that lists the variation of commonly measured clinical laboratory tests [326]. However, most of the data for this comes from trials with small numbers of participants, often measured in ideal conditions. A recent systematic review of the within-individual variation of CRP (chapter three) found only 34 studies that reported this finding for CRP and 26 that reported it for hsCRP with a total of 92,031 subjects in all the studies combined (range 4 to 56,218). For CRP the median CV was 0.41 with a range of 0.11 to 0.89 and for hsCRP the median CV was 0.44 (range 0.27 to 0.76). However, many of these studies had small numbers of participants and/or small numbers of repeat measurements, with four out of five of the studies with more than 1000 patients only having two measurements of CRP. Little difference between subgroups based on health conditions was found in this paper but low numbers of studies, measurements and participants made it hard to draw conclusions.

This study aimed to use real-world data (i.e. data from clinical sources rather than experimental studies) from the IQVIA Medical Research Database (IMRD) to estimate total variation of HbA1c in patients with and without diabetes, to compare this to previously published data, and to consider the clinical implications of CRP variation in relation to diagnosis, monitoring and management of acute respiratory disease.

7.3 Aims

To describe the within-individual variation of CRP in a large cohort of patients in a database of routine general practice records, to describe any covariates

associated with the variation and to describe the probability of within-individual variation in CRP resulting in clinically important changes in CRP.

7.4 Methods

The methods for this cohort study were as described in Chapter 6 for the HbA1c cohort study, with exceptions noted below. The study was reported according to the STrengthening the Reporting of Observational studies in Epidemiology (STROBE) statement for observational studies.

7.4.1 Participants

All participants in the database were eligible for inclusion. Except for sensitivity measures, participants were excluded if they did not have at least four repeat measurements of CRP expressed in mg/L recorded. Participants were followed until the earliest date of either death, leaving the practice, stopping contributing to the database or the study end date.

7.4.2 Variable choice

The primary outcome was within-individual variation of CRP calculated as the within-individual coefficient of variation (CV_T). For subgroup analysis, the dates of the CRP tests and additional variables including patient sociodemographic characteristics, lifestyle factors and diagnoses were extracted (see table 7.1). Variables were selected as those I hypothesised most likely to have an effect on

variation based on clinical plausibility or that enabled calculation of variation (e.g. date of measurement). Variables for sensitivity analysis were chosen to assess factors which might affect the robustness of the results such as statistical methods, number of repeat measurements and unit of measurement.

Table 7.1 Variables extracted from database.

Variables extracted	
Test characteristics	CRP result
	Date of measurement
	Unit of measurement
Patient characteristics	
Sociodemographic	Age
	Sex
	Ethnicity
	Townsend Deprivation Score
	BMI
	Geographical region
Lifestyle factors	Smoker status
	Alcohol consumption status
Diagnoses and comorbidities	Diabetes
	Hypertension
	Hyperthyroidism
	Hypothyroidism
	Ischaemic heart disease
	Heart failure
	Ischaemic stroke
	Haemorrhagic stroke
	All cancers

7.4.3 Data cleaning

CRP measurements are routinely expressed in mg/l, so other units of measurement were excluded. CRP must be 0 or above, therefore all negative values were excluded. From examination of a frequency distribution of the CRP values within the dataset, and published ranges of CRP values, values greater than 1000 were excluded as implausible [347].

7.4.4 Statistical methods

Since the CRP values were highly skewed, log-transformed data was used when calculating CVs [348]. The mixed command in Stata was then used to calculate within-individual variance using a random effects model. 95% confidence intervals for CVs were derived from the model in Stata. For consistency with other research, CV_T was expressed as a decimal fraction. It was considered improbable that participants would have identical CRP results on four or more different occasions and that these were likely duplicated results, so these participants were excluded. If more than one measurement was performed on the same day, only the first result was included on the basis that others recorded for the same day were likely to be duplications or errors.

There is little published data on minimal clinically important difference (MCID), and this may be due to the highly skewed nature of CRP results, with vastly different results encountered in different disease processes and with differing disease severity. Results may be interpreted and perceived differently in different diseases. For example a 5 mg/L change in CRP level may be perceived very differently by a patient with rheumatoid arthritis and an initial CRP of 10mg/L compared to a patient with an initial CRP of 150mg/L with bacterial pneumonia. The probability of getting a second result of 10mg/L (chosen arbitrarily as there is no consensus MCID) higher or lower than the initial result was calculated by counting the number of participants in which every combination of first and second results was achieved (e.g. 1st=0-10, 2nd=50-60; 1st=10-20, 2nd=60-70 etc). The probability of each combination of first and second results was then calculated by dividing the number of participants with that combination by the total number of participants who had the first result.

Statistical code for Stata is available from the author on request.

7.5 Results

Participants

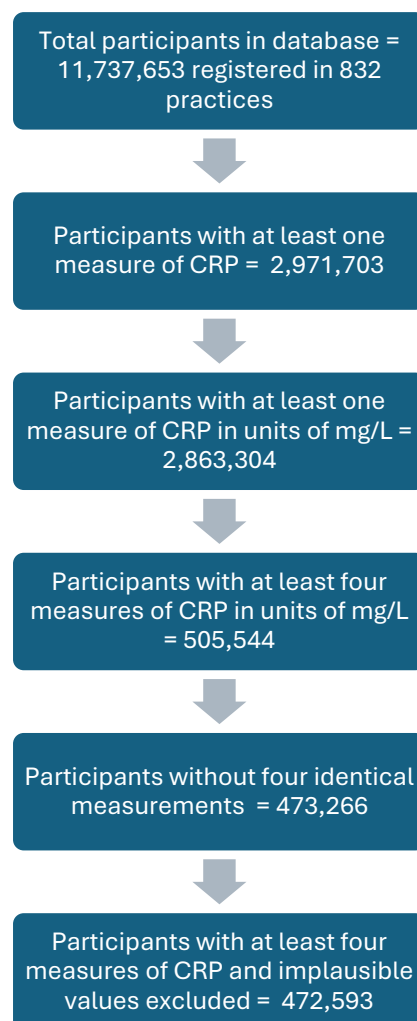


Figure 7.1 Participant selection algorithm

11,737,653 participants registered in 832 practices were available in the database for analysis. Figure 7.1 shows the flowchart of participant selection. 472,593 participants had at least four measures of CRP recorded as mg/L between 0 and 1000. The median number of measurements per individual was 10 (IQR 6-32).

7.5.1 Descriptive data and outcome data

The overall coefficient of variation CRP was 1.148 (95% CI 1.147 to 1.149), which is consistent with a very high level of within-individual measured variation.

Table 7.2 shows the participant characteristics and main univariable subgroup analyses. 472,593 participants were included in the study, with a median age at first measurement of 68 (IQR 54 to 80), 34.6% being male. Missing ethnicity was recorded for 239,643/472,593 (50.7%) participants and “Missing” Townsend deprivation score for 66,651/472,593 (14.1%) participants. The median number of years between first and last measurement in the database was 7 (IQR 4 to 10).

Table 7.2 Sociodemographic data and main univariable subgroup analyses.

		Median (mg/L)	CV(95% CI)	N(%)
Whole population		5.0	1.148 (1.147 to 1.149)	472,593 (100)
Deprivation quintile (1=least deprived 5=most deprived)	1	5.0	1.184 (1.181 to 1.187)	97,819 (20.7)
	2	5.0	1.151 (1.148 to 1.154)	92,516 (19.6)
	3	5.0	1.146 (1.144 to 1.149)	90,656 (19.2)
	4	5.0	1.127 (1.124 to 1.130)	74,503 (15.8)
	5	5.4	1.113 (1.109 to 1.117)	50,448 (10.7)
	Missing	5.0	1.134 (1.130 to 1.137)	66,651 (14.1)
Ethnicity	Black	5.0	1.065 (1.050 to 1.080)	3,603 (0.8)
	Mixed	4.0	1.045 (1.021 to 1.069)	1,103 (0.2)
	Other	4.0	1.084 (1.069 to 1.099)	3,303 (0.7)
	South Asian	4.0	1.000 (0.992 to 1.008)	8,797 (1.9)
	White	5.0	1.136 (1.135 to 1.138)	216,144 (45.7)
	Missing	5.0	1.164 (1.162 to 1.165)	239,643 (50.7)
Sex	Male	5.0	1.080 (1.079 to 1.081)	163,647 (34.6)
	Female	5.0	1.283 (1.281 to 1.286)	308,946 (65.4)
Age (years)	≤10	3.0	1.739 (1.591 to 1.911)	125 (0.0)
	>10 to 20	3.0	1.452 (1.425 to 1.480)	2,390 (0.5)
	>20 to 30	4.0	1.311 (1.301 to 1.321)	15,329 (3.2)
	>30 to 40	4.0	1.150 (1.145 to 1.156)	29,887 (6.3)
	>40 to 50	4.0	1.052 (1.049 to 1.056)	47,907 (10.1)
	>50 to 60	5.0	1.010 (1.008 to 1.013)	76,827 (16.3)
	>60 to 70	5.0	1.062 (1.060 to 1.065)	85,470 (18.1)
	>70 to 80	5.0	1.156 (1.153 to 1.159)	97,108 (20.5)
	>80 to 90	6.0	1.271 (1.267 to 1.274)	79,738 (16.9)
	90 to 100	7.0	1.390 (1.384 to 1.397)	34,431 (7.3)
Median CRP (mg/L)	0 to <1	0.6	1.479 (1.468 to 1.490)	14,692 (3.11)
	1 to < 2	1.0	1.149 (1.145 to 1.152)	61,503 (13.01)
	2 to <3	2.0	1.188 (1.184 to 1.192)	49,370 (10.45)
	3 to < 4	3.0	1.083 (1.080 to 1.086)	50,904 (10.77)
	4 to <5	4.0	1.063 (1.060 to 1.066)	53,351 (11.29)
	5 to <6	5.0	0.981 (0.978 to 0.983)	59,021 (12.49)
	6 to < 7	6.0	1.075 (1.070 to 1.079)	23,709 (5.02)
	7 to < 8	7.0	1.071 (1.066 to 1.076)	19,408 (4.11)
	8 to < 9	8.0	1.041 (1.036 to 1.046)	18,285 (3.87)
	9 to < 10	9.0	1.054 (1.048 to 1.060)	13,184 (2.79)
	10 to <20	13.0	1.088 (1.086 to 1.091)	65,047 (13.76)
	20 to < 30	23.4	1.220 (1.214 to 1.226)	20,035 (4.24)
	30 to < 40	34.0	1.349 (1.338 to 1.360)	8,942 (1.89)
	40 to < 50	44.0	1.463 (1.446 to 1.480)	5,047 (1.07)
	50 to < 60	54.0	1.612 (1.585 to 1.639)	2,995 (0.63)
	60 to < 70	64.0	1.621 (1.587 to 1.655)	1,917 (0.41)
	70 to < 80	74.0	1.746 (1.699 to 1.794)	1,350 (0.29)
	80 to < 90	84.0	1.832 (1.773 to 1.894)	946 (0.20)
	90 to < 100	94.0	1.942 (1.861 to 2.028)	665 (0.14)
	100 to < 200	122.0	2.100 (2.046 to 2.157)	2,036 (0.43)
	200 to < 300	223.0	2.150 (1.960 to 2.372)	170 (0.04)

7.5.1.1 Subgroup analyses

Subgroup analysis of CV_T was also performed by sex, age, ethnicity, geographical region, Townsend deprivation score, BMI, presence of comorbidities,

alcohol consumption and smoking status (see Table 7.2 and Appendix 7.1 figs 17.1 to 17.9). Higher CV_Ts were seen in females than males, in white ethnicities compared to other ethnicities and in higher deprivation areas compared to lower deprivation areas.

Lower BMI was associated with a higher variation than higher BMIs. A U-shaped relationship with age was seen with a nadir of CV_T around 50 to 60 years of age. Compared to no comorbidity, patients with chronic obstructive pulmonary disease (COPD), chronic kidney disease (CKD) stage 3 to 5, heart failure, ischaemic heart disease, ischaemic stroke, ischaemic heart disease or cancer had a higher CV_T. Higher mean and median CRP levels were associated with higher variation.

Median CRP level, deprivation score, ethnicity, sex and age had larger effects on CV, so these were chosen for multi-subgroup analysis. The results of this are presented in a heat map in fig 7.2. It can be seen from this that of these subgroups, median CRP level seems to have the most marked effect on CV.

	Sex Age Ethnicity	Male				Female			
		20-60		>60		20-60		>60	
		White	Non-white	White	Non-white	White	Non-white	White	Non-white
CRP	Deprivation								
0 to 20	1 to 3	1.195	1.216	1.266	1.288	1.012	1.004	1.087	1.106
	4 to 5	1.160	1.138	1.204	1.276	0.963	0.975	1.068	1.108
>20 to 100	1 to 3	1.612	1.776	1.548	1.522	1.247	1.151	1.373	1.358
	4 to 5	1.518	1.460	1.435	1.465	1.053	1.098	1.235	1.296
>100	1 to 3	2.639	2.715	2.035	2.076	3.245	3.120	2.232	1.991
	4 to 5	2.360	2.429	1.866	1.977	2.033	2.445	1.768	1.960

Figure 7.2 Heat map showing the coefficient of variation for multivariable analysis according to median CRP result, sex, age, ethnicity, deprivation and BMI.

Age is in years; deprivation score is Townsend deprivation score.

Figure 7.3 shows the CV by patient median CRP level, which demonstrates a marked increase in variation with patient increasing patient median.

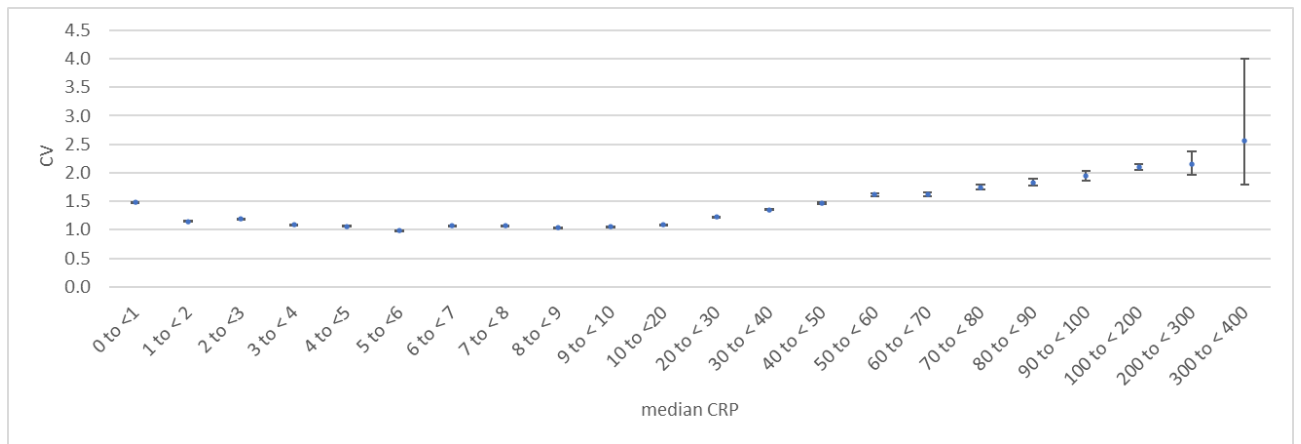


Figure 7.3 Coefficient of variation by patient median with 95% confidence intervals

																					Total probability of this difference	
	second result																				Difference	
first result	0 to 10	10 to < 20	20 to < 30	30 to < 40	40 to < 50	50 to < 60	60 to < 70	70 to < 80	80 to < 90	90 to < 100	100 to < 110	110 to < 120	120 to < 130	130 to < 140	140 to < 150	150 to < 160	160 to < 170	170 to < 180	180 to < 190	190 to < 200	10 or less	10 or more
0 to 10	81.70	8.47	2.82	1.57	1.04	0.76	0.57	0.48	0.35	0.30	0.26	0.21	0.18	0.14	0.13	0.11	0.11	0.09	0.07	0.06		17.72
10 to < 20	47.83	32.53	8.10	3.43	1.87	1.20	0.94	0.55	0.54	0.38	0.35	0.31	0.21	0.18	0.15	0.14	0.11	0.11	0.09	0.10	47.83	18.75
20 to < 30	41.86	24.14	15.45	6.70	3.26	1.95	1.37	0.90	0.67	0.50	0.56	0.34	0.31	0.24	0.24	0.17	0.16	0.07	0.09	0.09	66.00	17.61
30 to < 40	41.45	18.64	13.17	9.59	5.09	3.13	2.08	1.27	1.02	0.85	0.69	0.43	0.28	0.30	0.30	0.22	0.19	0.13	0.10	0.09	73.26	16.18
40 to < 50	41.95	16.63	10.48	8.27	7.44	4.18	2.75	1.64	1.26	0.88	0.75	0.43	0.39	0.42	0.31	0.30	0.24	0.16	0.12	0.12	77.34	13.96
50 to < 60	41.36	16.33	9.22	7.65	5.89	5.55	3.95	2.59	1.66	1.12	0.77	0.57	0.49	0.47	0.32	0.37	0.20	0.20	0.15	0.05	80.44	12.90
60 to < 70	40.18	16.05	9.58	6.70	5.52	4.90	4.06	2.82	2.15	1.57	1.24	0.93	0.67	0.60	0.49	0.42	0.31	0.20	0.18	0.16	82.93	11.73
70 to < 80	40.06	15.34	8.79	5.71	5.06	4.80	4.12	3.95	2.82	1.84	1.64	0.88	0.79	0.79	0.59	0.37	0.51	0.17	0.23	0.31	83.87	10.93
80 to < 90	39.43	15.23	8.83	6.14	4.47	4.54	4.00	3.71	3.82	2.33	1.42	1.05	0.87	0.58	0.80	0.33	0.33	0.40	0.25	0.15	86.34	8.50
90 to < 100	38.26	13.94	9.03	6.89	4.79	3.11	3.15	3.70	3.15	3.61	2.56	1.43	1.26	0.92	0.67	0.55	0.50	0.34	0.42	0.25	86.01	8.90
100 to < 110	35.20	16.63	8.68	6.44	4.88	4.05	3.22	3.51	2.88	2.93	2.78	1.56	1.12	1.51	0.83	0.39	0.49	0.88	0.24	0.39	88.40	7.41
110 to < 120	33.98	15.45	8.05	5.68	4.68	5.45	4.14	2.66	2.96	2.43	2.96	3.43	1.95	1.30	1.01	0.83	0.77	0.30	0.59	0.24	88.45	6.99
120 to < 130	32.82	16.09	9.42	5.34	5.27	4.57	3.30	2.60	2.81	2.18	2.39	2.18	1.90	1.83	0.70	1.48	0.91	0.98	0.42	0.49	88.97	6.82
130 to < 140	35.15	13.56	8.03	4.84	5.35	3.80	2.94	3.11	2.68	2.33	2.85	2.25	2.25	3.02	1.73	0.95	1.04	0.86	0.78	0.60	89.12	5.96
140 to < 150	33.13	13.05	8.56	4.89	4.18	4.38	4.79	3.77	2.55	1.73	2.55	1.53	1.12	2.14	2.45	1.73	1.94	0.51	0.71	0.51	88.38	5.40
150 to < 160	30.07	15.34	7.85	6.40	4.23	3.38	3.50	3.62	2.78	2.29	2.54	2.66	2.54	1.93	2.05	2.29	1.45	0.97	0.60	0.97	91.18	3.99
160 to < 170	30.98	14.86	8.19	6.68	4.41	3.40	2.52	3.02	2.52	2.77	1.89	1.64	1.39	2.52	1.39	1.26	2.14	1.51	1.76	0.63	89.42	3.90
170 to < 180	31.53	16.38	7.73	4.95	4.64	3.55	2.78	2.63	2.94	1.85	2.94	3.55	1.39	2.32	1.24	1.08	0.93	2.01	0.77	0.93	92.43	1.70
180 to < 190	28.91	13.90	9.58	6.07	4.79	4.31	3.83	2.40	2.24	2.40	1.60	1.92	1.28	2.08	1.28	1.76	1.44	1.44	2.72	1.12	91.21	1.12
190 to < 200	30.24	14.23	8.50	6.72	4.94	3.56	3.75	1.58	3.75	2.77	2.57	1.38	1.78	1.19	1.78	1.98	1.58	0.79	1.19	1.58	94.27	

Figure 7.4 Probability of second result being 10mg/L different than first result.

The last two columns show the probability of the second result being at least 10 or more mg/L different than the first score.

Fig 7.4 shows a heat map with data on the probability of the result of a second CRP test given the result of an initial CRP test.

7.5.1.2 *Sensitivity analyses*

CV_T was calculated by year, quarter, number of measurements, patient mean CRP, median days between measurements, alternative methods of calculating CV (linear regression on non-log adjusted data, and a crude arithmetic method of calculating CV), and CV by patient mean on non-log adjusted data (see Appendix 7.2 figs 17.10 to 17.15). Little difference in CV was seen when stratified by year or quarter suggesting no trend over time in variation and no seasonal effect. An increased number of measurements was associated with a lower CV, perhaps because more measurements smoothed out spikes in CRP from transient illness. However, CV was markedly higher if the median time between measurements was less than about one month. There was little difference in CV when calculated by crude arithmetic methods and log-adjusted linear regression.

7.6 Discussion

7.6.1 Key results

This is the largest study of variation of CRP to date, with 472,593 participants, which is nearly five times more than the number of participants in all 60 published studies reporting CRP variation put together described in a recent systematic review (See Chapter 3).

The overall coefficient of variation for CRP was 1.148 (95% CI 1.147 to 1.149). This is much higher than the median reported for previous studies of 0.41 (range

0.11 to 0.89), see Chapter 3. The probability of a second measurement differing markedly from a first measurement is high as shown in figure 7.4, although this finding must be interpreted in conjunction with the fact that CRP has a short half-life of around 19 hours after the inflammatory trigger is removed [349]. Furthermore, the data does not distinguish between situations where something has materially changed e.g. a change of medication or resolution of disease, and where nothing has changed.

Subgroup analyses showed that age, sex, ethnicity, BMI, and presence of certain comorbidities had an effect on CV_T of CRP, with younger and older participants, female participants and participants with lower BMIs having higher CV_T s, likely due to a higher degree of pathological variation in these groups. Note that there may be some interaction between subgroups, with for example age, smoking, deprivation and BMI being associated with each other and with comorbidities. The median age was relatively high at 68 years which might be associated with a higher degree of pathological variation.

However, as the heat map of multiple subgroup analysis (fig 7.2) and the graph of CV_T versus median patient CRP (fig 7.3) show, the effect of median patient CRP has a much larger effect than other subgroup analyses, with the CV_T increasing markedly with increasing median patient CRP after a value of around 10mg/L. This is important because patients with more severe disease are likely to have a higher variation of results, making treatment and diagnosis decisions more difficult. For example at the cut-off point recommended by NICE for giving immediate antibiotics to patients with respiratory disease, 100mg/L, the CV is about 70% higher than the overall population CV (1.94 versus 1.15).

7.6.2 Strengths and limitations

The major strength of this study is the large size of the dataset which gives more accurate overall CV and allows the investigation of subgroups. It is also important that the data is real-world data taken from the clinical records of UK primary care practices, which are used for clinical decision-making, and not measurements taken under ideal or experimental conditions. The Biological Variation Data Critical Appraisal Checklist (BIVAC) recommends biological variation studies are performed on patients in steady state, with “preanalytical procedures described and standardized to minimize preanalytical variation” [154]. This is important when using biological variation data to set analytical performance specifications, but is less relevant to clinical decision-making. Thus the real-world nature of the data in this study can be considered a weakness from the point of view of setting analytical standards, but it is also a strength in that real-world data is the data used to make clinical decisions and it includes every component of variation (pre-analytical, analytical, biological and pathological).

Assumptions were made for the repeat results calculation, namely that the change between first and second result is representative of changes in the result in subsequent calculations, which may not be the case.

Another weakness with the data is that it is hard to account for confounding by indication, for example patients with greater disease severity are likely to have more frequent measurements. Nevertheless, the large disparity between this study and previously published results and the large number of participants help to compensate for this problem.

A further potential weakness with the data is error associated with data entry and recording into the clinical databases (although the electronic transmission of lab results directly into databases is likely to minimise this). Data cleaning prior to analysis also partially compensates for this. For example implausible values were removed (taking the cut-off points from analysis of the IMRD data itself and published data [347]), duplicate results on the same day were removed, participants with less than four repeated measurements were eliminated, and any patients with four or more identical results were removed, since this is a highly unlikely finding.

Inclusion of the entire dataset meant the only potential bias introduced into the results (that was not already present in the database) could come from data cleaning (e.g. exclusion of implausible values) and the specification that at least four measurements must be included. However, this latter specification, although aimed at increasing the accuracy of the variation calculation, could introduce confounding by indication since patients with more frequent tests might have more severe/less stable disease and hence a higher degree of pathological variation. Furthermore, it has been shown that the ordering of any test by a clinician increases the risk of the disease which the test is investigating, since the presence of a test in a patient record suggests there was a suspicion that patient was at risk of the disease. This could cause some degree of selection bias when compared to results obtained from healthy populations in controlled biological variation studies [338].

Finally the nature of CRP as an acute phase protein with a short-half life has implications for assessment of CRP variation over a long term. Fig 17.14 shows that patients with a very short median interval between measurements have markedly higher variation, which is consistent with the change in CRP levels with acute disease progression and resolution and treatment. Nevertheless, the longer term

within-individual variation (where the median interval between measurements is more than about a month) is quite consistent. Furthermore, CRP is used in the monitoring of chronic disease, e.g. as a component of the DAS-28 instrument for monitoring rheumatoid arthritis [350] and so the long-term variation of CRP is clinically important.

7.6.3 Interpretation and generalisability

This study shows that the variation of CRP in this population of primary care patients in the UK is much higher than results for variation previously reported in the literature. This may be because the participants were less likely to be in a steady state, and the pre-analytical testing conditions were not rigorously controlled. However, as discussed, this is the data on which clinical decisions are based.

The findings of this study have clinical implications for diagnosis and monitoring. Not taking within-individual variation into account when making a diagnosis could lead to false positives and hence inappropriate treatment of for example rheumatoid arthritis with immunosuppressives leading to possible adverse effects, or bacterial pneumonia with antibiotics which could contribute to antimicrobial resistance. False negatives in diagnosis however could lead to delayed or missed opportunities for treatment, leading to poorer patient outcomes. Similarly, when using CRP for monitoring, for example of rheumatoid arthritis, a measured CRP value that is much higher than the real value may lead to inappropriate treatments with immunosuppressive drugs, increasing the risk of adverse effects, while a measured CRP value that is much lower than the real value may lead to treatment being withheld, causing patient morbidity and potential disease progression.

NICE guidelines recommend that CRP results are used to guide antibiotic treatment for acute lower respiratory tract disease. However, the high within-individual variation of CRP, especially at the higher values, could mean that a measured value is not reflective of the patient's true mean, so antibiotic treatment is inappropriately withheld. Similarly, AGA guidelines on the management of Crohn's disease recommend CRP is used to guide whether endoscopic investigation is required, and the high variation of CRP could mean that a patient is classified incorrectly as requiring or not requiring endoscopy [346].

The chart showing the probability of a second CRP result given the first result can be used to give some estimation of the extent of the problem. For example a patient with a first CRP measurement of 60-70mg/L has an 82.9% chance of the second result being 10mg/L lower than the first result and a 11.7% chance of the second result being 10mg/L or more higher than the first result. The high chance of the second result being lower than the first may not only reflect variation, but some degree of regression to the mean, since diseases for which CRP are part of the diagnostic and monitoring process are often acute, and CRP has a short half-life. The fact that over 10% of patients show an important increase in CRP despite this fact helps confirm the high variation of CRP.

The CV calculated in this study can also be used to estimate the chance of getting a particular measured result given the true mean of a patient. For example, if we take a hypothetical patient with a true CRP of 60 mg/L (a value in the middle of the equivocal group for recommending an antibiotic for respiratory tract infection), the CV from the data in this study is approximately 2.4. This suggests a probability of 39% that the measured result will be 20mg/L or lower, which would mean antibiotics were not prescribed, and a probability of 39% that the measured result will be

100mg/L or higher, meaning that immediate antibiotics are recommended. Thus the high variation of CRP means that it is highly likely that single measurements will incorrectly categorise a patient when making treatment decisions.

Clinicians should be aware of the high variation of CRP and the high probability of incorrect inferences from CRP results that can result from this. They need to be aware of the possibility for misdiagnosis and inappropriate management of inflammatory and infectious conditions when relying on single measures of CRP, (although even taking multiple measurements does not completely eliminate this problem) and authors of clinical guidelines should take variation into account when making diagnosis, monitoring and treatment recommendations.

Note that some of the underlying conditions for which CRP is part of the diagnostic or monitoring pathway are more variable than other conditions (e.g. bacterial pneumonia versus rheumatoid arthritis), and also more variable than other underlying conditions for which the tests examined in this thesis are performed. That is to say that the pathological variation in those conditions is higher in some conditions than others. Examples of conditions with a high degree of pathological variation include bacterial pneumonia and other acute illnesses, while those with a lower degree of pathological variation include chronic diseases such as rheumatoid arthritis and Crohn's disease.

Further research that addresses confounding by a change in frequency or timing of testing with increased disease severity or variation would be useful. It would be useful to perform a similar study restricted to patients with only long-term inflammatory diseases, to reduce the issue of confounding by fluctuation of the natural disease course. Studies into how variation in this population affects clinical

outcomes such as mortality and whether reducing variation with improved testing and management protocols improves outcomes is also important. Other options for future research would be to work out the variation when patients are in a stable condition with respect to medication and disease.

7.7 Conclusions

This study suggests that the total measured within-individual variation of CRP in real-world situations is much higher than previously reported in the literature. Within-individual variation of CRP increased with median CRP after values for CRP of around 10mg/L.

The next chapter uses the IMRD database to estimate the short-term (<6 months) variation of the spirometry measures FEV1 and FVC, to estimate the within-individual CV in a non-laboratory test.

8 Chapter eight: Within-individual variation of spirometry measurements in primary care: a retrospective cohort study

8.1 Abstract

Background

FEV1 and FVC are two of the most common measurements in spirometry, which is used for the diagnosis and monitoring of respiratory disease. Variation in within-individual measured FEV1 and FVC may affect their clinical utility but estimates of this are based on limited and often poor-quality data.

Methods

Data on FEV1 and FVC results and sociodemographic, lifestyle and comorbidity covariates were extracted from the IMRD database using the DEXTER tool. A minimum of four measurements in the same individual within a six-month time period from the first measurement was the only inclusion criterion. Analysis was restricted to measurements within six months of the first measurement in order to reduce the effect of longitudinal decline in lung-function on the estimates. Within-individual measured variation was calculated as a coefficient of variation (CV) using a linear regression random effects model for the whole population and various subgroups.

Results

4,412 participants for FEV1 and 3,567 participants for FVC were included in the main study.

The overall coefficient of variation for FEV1 was 0.224 (95% CI 0.221 to 0.228) and for FVC was 0.152 (0.150 to 0.155). This is much higher than seen in a previous systematic review of spirometry variation which reported a median CV for FEV1 of 0.034 and for FVC of 0.031. CV increased as mean patient FEV1 and FVC decreased, i.e. variation increased as disease severity worsened.

Strengths and limitations

The large number of participants and the real-world nature of the results are important strengths of this study. Weaknesses included difficulties accounting for confounding by indication.

Conclusions

Estimated within-individual variation in this analysis of real-world data is very high and is higher than previously reported. Variation increases with decreasing patient mean i.e. with more severe disease status. This has important implications for the diagnosis, monitoring and clinical decision-making for diabetes.

8.2 Introduction

Spirometry is recommended for the diagnosis and management of asthma and chronic obstructive pulmonary disease (COPD) [351]. In England it is commonly delivered in primary care. Concerns have been expressed over the technical quality and reliability of spirometry in this setting, although the real-world usefulness of the test is still considered to be high [352]. Forced Expiratory Volume in one second

(FEV1) and Forced Vital Capacity (FVC) are two of the most common measurands assessed by spirometry. These are usually reported in absolute values of volume (litres or mLs) or in relative values (percentage of predicted value derived from reference tables that take into account height, age, sex and ethnicity) [178, 353]. Individual variation may differ between different subgroups of patients e.g. with age, health status, sex or ethnicity.

Current NICE guidelines suggest the presence of mild airflow obstruction is indicated by a FEV1 value of between 50% and 79% of the predicted value, with less than 30% indicating severe airflow obstruction (see table 8.1) [354].

Table 8.1 NICE guidelines for categorising severity of airflow obstruction.

FEV1 % predicted	NICE guideline CG12 (2004) severity of airflow obstruction
≥ 80%	Not categorised
50–79%	Mild
30–49%	Moderate
< 30%	Severe

Post-bronchodilator reversibility can be used to help confirm the diagnosis of COPD. However, NICE guidelines do not routinely recommend the use of this test in diagnosis since repeated FEV1 measurements can show spontaneous fluctuation, the results performed on different occasions may not be reproducible, and the definition of the value of an important change is arbitrary [354]. NICE guidelines also state that repeated reversibility testing is not recommended

These guidelines acknowledge that spirometry testing, like all biological measurements, vary within an individual over time. This within-individual variation is broadly composed of three parts. Pre-analytical variation is variation due to varying

factors prior to a measurement. For spirometry, this may include factors such as recent exercise, smoking, or consumption of alcohol or a large meal [355]. Analytical variation is the variation introduced by imprecision in the measuring process. In the case of spirometry this may relate to the technical quality of the spirometry testing. Biological variation is the within-individual variation, influenced partly by predictable factors such as season, time of day or monthly hormone cycles, but also to a large extent by chance. The sum of these three types of variation gives the total within-individual variation [15]. The higher the within-individual variation, the lower the probability that a single measurement is reflective of the true mean. In this study, it is not possible to separate analytical from biological variation, so within-individual coefficient of variations (CV_T) is taken to refer to total within-individual variation i.e. the sum of biological and analytical variation.

As the NICE guidelines imply, it is important to be aware of the within-individual variation of spirometry testing to quantify the probability of errors in diagnosis and management of respiratory illness. However, a systematic review of within-individual variation of spirometry found that most studies published on this subject had small number of participants and/or small numbers of repeat measurements (Chapter four). 58.7% of the studies reported between-day variation, with the rest reporting within-day variation or it being unclear which they were reporting. It is acknowledged that within-session variation of spirometry is high, and American Thoracic Society guidelines recommend repeated tests are performed in order to achieve three acceptable measurements of FEV₁ and FVC, with repeatability recommendations based on the difference between the largest and next largest measurements [356]. However, less guidance is available about how often

tests should be repeated on different days to overcome the problem of longer term within-individual variation.

Spirometry measures can decline over time with age and disease progression. For example, patients with COPD can show a mean decline in FEV1 of 33-69mL/year with some patients showing as rapid a decline as 200mL/year [357]. Variation over a long time-period, therefore, includes both short-term within-individual variation and a long-term decline. Restricting analyses to measurements performed within a six-month period therefore will provide an estimate of short-term variation.

This study aimed to use real-world data from the IQVIA Medical Research Database (IMRD) to estimate total variation of FEV1 and FVC, to compare this to previously published data, and to consider the clinical implications of variation of spirometry in relation to diagnosis, monitoring and management of COPD and asthma.

8.3 Aims

To describe the within-individual variation of FEV1 and FVC in a large cohort of patients in a database of routine general practice records, and to describe any patient characteristics or other factors associated with the variation.

8.4 Methods

The methods for this cohort study were as described in Chapter 6 for the HbA1c cohort study, except as noted below. The study was reported according to the STrengthening the Reporting of Observational studies in Epidemiology (STROBE) statement for observational studies.

8.4.1 Participants

All participants in the database were eligible for inclusion. The primary analysis used all measurements within a six-month period, for which there were at least four measurements in order to estimate short-term variation rather than include variation due to the long-term decline with age or disease progression. Participants were excluded if they did not have at least four recorded measurements of FEV1 or FVC recorded in a six-month period. All measurements more than six months after the first measurement were excluded to ensure decline with age or progression of disease did not lead to erroneous calculations of variation (except for a sensitivity analysis that looked at measurements more than six months after the first measurement). Unless otherwise stated, all calculations in this study were performed on these short-term measurements.

8.4.2 Variable choice

The co-primary outcomes were within-individual variation of FEV1 and FVC as demonstrated by the within-individual coefficient of variation (CV_T). Other variables included in analyses are shown in table 8.2. Variables were selected as those I hypothesised most likely to have an effect on variation based on clinical plausibility

or that enabled calculation of variation (e.g. date of measurement). The date of each FEV1 or FVC test and the date of diagnosis where applicable were also extracted.

Table 8.2 Variables extracted from database.

Variables extracted	
Test characteristics	
	Date of measurement
	Unit of measurement
Patient characteristics	
Sociodemographic	Age
	Sex
	Ethnicity
	Townsend Deprivation Score
	BMI
	Geographical region
Lifestyle factors	Smoker status
	Alcohol consumption status
Diagnoses and comorbidities	Prediabetes
	Type one diabetes mellitus
	Type two diabetes mellitus
	Diabetic medications
	Hypertension
	Hyperthyroidism
	Hypothyroidism
	Ischaemic heart disease
	Heart failure
	Ischaemic stroke
	Haemorrhagic stroke
	All cancers

8.4.3 Bias and study size

The study size included the entire database where participants had four or more measurements of FEV1 and FVC within six months of the first measurement (see figure 8.1).

8.4.4 Data cleaning

Most of the spirometry measurements were reported in litres so those expressed as a percent predicted were excluded from the main analysis since the units are not interchangeable with regards to calculation of coefficient of variation (but values recorded as percent predicted were used for a sensitivity analysis). Values of FEV1 or FVC more than 10L were excluded as being implausible based on published data e.g. [358] and analysis of the IMRD data.

8.4.5 Statistical methods

Since FEV1 and FVC values are often expressed as a percentage, it was elected to express the CV_T as a decimal fraction in this study. Participants with identical FEV1 or FVC results on at least four different occasions were considered improbable and likely duplicated results, so were excluded. Univariable subgroup analyses were performed by age, sex, ethnicity, BMI, Townsend neighbourhood deprivation score quintiles, Townsend neighbourhood deprivation score quintiles, (which give a score from 1 to 5 estimating deprivation in a local area, with 1 being the least deprived and 5 being the most deprived [334]), smoking and alcohol status, geographical location, mean and median FEV1 or FVC result and the presence of COPD, asthma and other comorbidities. Based on the variation calculated in the univariable analysis and the likely importance of factors in respiratory disease (e.g. smoking status), several subgroups were selected to be combined into a multiple subgroup analysis to describe how within-individual variation of spirometry measurements differs between subgroups of individuals. The probability of getting a second result of 0.5L higher or lower than the initial result was calculated by using the dataset to evaluate the

number the number of participants in which every combination of first and second results was achieved. The probability of each combination of first and second results was then calculated by dividing the number of participants with each combination by the total number of participants who had the first result.

Age and sex were complete throughout the data set. Participants with missing FEV1 and FVC results were not eligible. Where other comorbidities were not reported, they were assumed to be absent. Participants with missing data for other characteristics such as ethnicity, BMI or deprivation score were included in the main analysis but excluded from the subgroup analysis where these variables were used.

Various sensitivity analyses were performed to assess bias due to factors which might affect the robustness of the results such as statistical methods, number of repeat measurements and unit of measurement. An analysis was performed by using all the measurements in the database regardless of timing, i.e. including those measurements that were more than six months after the first measurement. For a sensitivity analysis based on year, participants were divided into calendar years to look for a trend over time. A similar analysis was performed for quarter where measurements in a particular quarter were grouped to look for seasonal trends. A sensitivity analysis was performed where participants were grouped based on reported unit (L v percentage predicted). Other sensitivity analyses were performed where participants were grouped according to the number of measurements of each participants, days between measurements and using various cut-off points to test the effect of excluding implausible values. An analysis using all the measurements in the database rather than just those within six months of each other was carried out. A simple arithmetic method of calculating CV_T was also performed in which the

individual standard deviation and mean were used to calculate an individual CV_T , which was then averaged using the mean.

8.5 Results

Participants

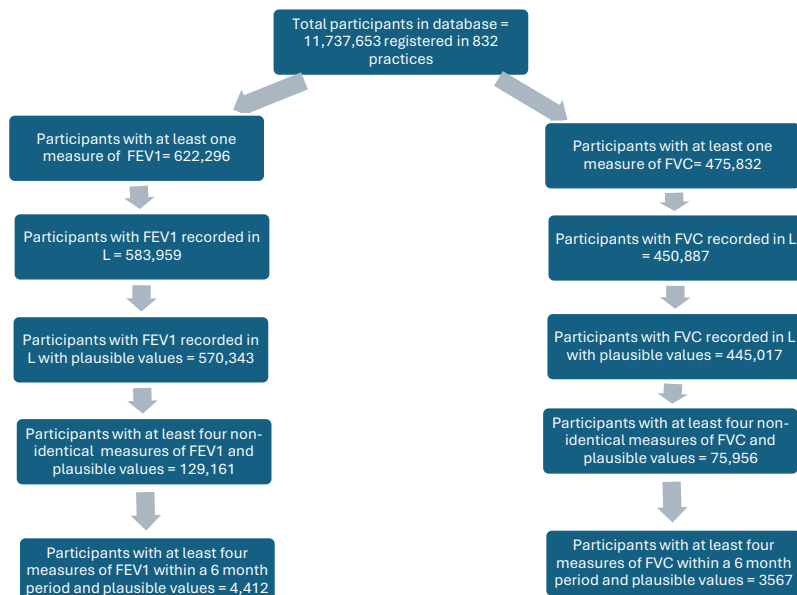


Figure 8.1 Participant selection algorithm

The database included 11,737,653 participants registered in 832 practices. 622,296 participants had at least one FEV1 and 475,832 had at least one FVC. Figure 8.1 shows the flowchart of participant selection. After exclusion of measures not recorded in litres and implausible values (not between 0 and 10 L) 4,412 participants had at least four measures of FEV1 within a six-month period and 3,567 participants

had at least four measures of FVC within a six-month period. The median number of measurements per individual was 8 (IQR 5-11) for FEV1 and 7 (IQR 5-10) for FVC.

8.5.1 Descriptive data and outcome data

The overall coefficient of variation for FEV1 was 0.224 (95% CI 0.221 to 0.228) with a mean FEV1 of 2.03 L. The overall coefficient of variation for FVC was 0.152 (95% CI 0.150 to 0.155) with a mean FVC of 3.03 L. Table 8.3 shows the participant characteristics and main univariable group analyses for FEV1 and FVC. Further univariable group analyses are found in tables 18.1, 18.2 and Figs 18.1 to 18.18). For FEV1, 4,412 participants were included in the study, with a median age at first measurement of 72 (IQR 56 to 72), 54.5% being male. For FVC, 3,567 participants were included in the study, with a median age at first measurement of 69 (IQR 57 to 79), 53.9% being male.

Table 8.3 Sociodemographic data and main univariable subgroup analyses for FEV1 and FVC.

x=<50 participants so CV not calculated

		FEV1 measurements					FVC measurements				
		N	%	Mean	CV	95%CI	N	%	Mean	CV	95%CI
All		4412	100	2.03	0.224	(0.221-0.228)	3,567	100	3.03	0.152	(0.15-0.155)
Sex	Male	2,404	54.5	2.29	0.229	(0.224-0.234)	1,921	53.9	3.51	0.149	(0.146-0.152)
	Female	2,008	45.5	1.71	0.206	(0.202-0.211)	1,646	46.1	2.47	0.152	(0.149-0.156)
Age	<=10	4	0.1	x	x	x	5	0.1	x	x	x
	>10 to 20	39	0.9	x	x	X	43	1.2	x	x	x
	>20 to 30	89	2	2.77	0.159	(0.142-0.176)	86	2.4	3.54	0.115	(0.103-0.127)
	>30 to 40	146	3.3	2.90	0.131	(0.121-0.141)	146	4.1	3.85	0.118	(0.109-0.127)
	>40 to 50	255	5.8	2.85	0.19	(0.18-0.201)	247	6.9	3.86	0.126	(0.119-0.133)
	>50 to 60	537	12.2	2.49	0.215	(0.206-0.223)	461	12.9	3.55	0.142	(0.136-0.148)
	>60 to 70	910	20.6	2.16	0.226	(0.219-0.233)	800	22.4	3.19	0.153	(0.148-0.158)
	>70 to 80	1,207	27.4	1.88	0.233	(0.226-0.239)	924	25.9	2.89	0.159	(0.154-0.163)
	>80 to 90	872	19.8	1.61	0.26	(0.251-0.269)	630	17.7	2.44	0.176	(0.169-0.182)
	90 to 100	331	7.5	1.39	0.247	(0.234-0.261)	211	5.9	2.13	0.175	(0.164-0.187)
Ethnicity	Black/Mixed/Others	51	1.2	2.28	0.165	(0.144-0.187)	44	1.2	x	x	x
	South Asian	74	1.7	2.18	0.191	(0.169-0.212)	53	1.5	2.76	0.225	(0.198-0.253)
	White	2,251	51	2.07	0.226	(0.221-0.231)	1,872	52.5	3.11	0.152	(0.149-0.156)
	Missing	2,031	46	1.96	0.225	(0.22-0.23)	1,591	44.6	2.96	0.148	(0.144-0.151)
BMI	<18.5	99	1.4	1.88	0.270	(0.238 - 0.302)	72	2.0	2.78	0.148	(0.129-0.166)

		FEV1 measurements					FVC measurements				
		N	%	Mean	CV	95%CI	N	%	Mean	CV	95%CI
	18.5 to <25	1,354	19.4	1.97	0.238	(0.231 - 0.245)	1,101	30.9	3.02	0.156	(0.151-0.16)
	25 to <30	1,285	18.4	2.08	0.222	(0.216 - 0.228)	976	27.4	3.07	0.15	(0.146-0.155)
	30 to <35	474	6.8	2.01	0.212	(0.203 - 0.222)	428	12	2.96	0.149	(0.142-0.156)
	>35	273	3.9	2.00	0.208	(0.196- 0.221)	208	5.8	2.78	0.178	(0.167-0.189)
	Missing	3,486	50.0	2.02	0.227	(0.223 - 0.231)	782	21.9	3.12	0.145	(0.140-0.151)
Smoker status	Current	1,651	37.4	1.97	0.253	(0.246-0.259)	1,289	36.1	3.03	0.158	(0.154-0.162)
	Ex	1,031	23.4	1.95	0.238	(0.23-0.246)	808	22.7	2.94	0.156	(0.151-0.161)
	Never	1,164	26.4	2.13	0.185	(0.179-0.191)	1003	28.1	3.04	0.146	(0.141-0.15)
	Missing	566	12.8			(-)	493	13.8			(-)
Respiratory status	Asthma	1,805	40.9	2.05	0.216	(0.21-0.221)	1,422	32.2	3.08	0.149	(0.145-0.153)
	COPD	1,514	34.3	1.72	0.291	(0.284-0.299)	1,087	24.6	2.8	0.175	(0.169-0.18)
	COPD & asthma	606	13.7	1.65	0.292	(0.28-0.305)	403	9.1	2.7	0.184	(0.175-0.193)

For FEV1, data was recorded as “Missing” for ethnicity in 2,031/4412 (46.0%) participants and “Missing” in Townsend deprivation score for 454/4412 (10.3%) participants. For FVC, data was recorded as “Missing” for ethnicity in 1591/3567 (44.6%) participants and “Missing” in Townsend deprivation score for 373/3567 (10.5%) participants.

8.5.1.1 Subgroup analyses

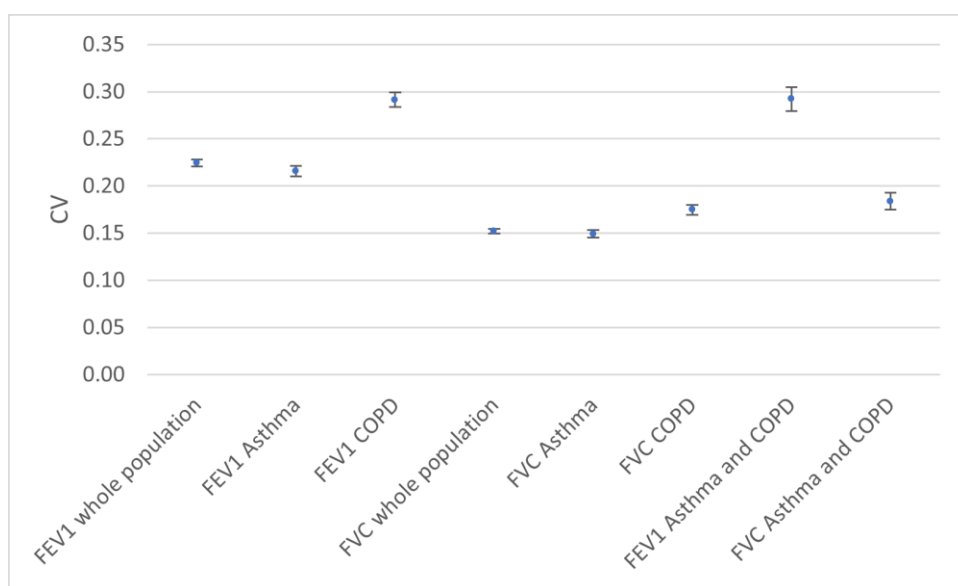


Figure 8.2 Graph of CVT of FEV1 and FVC by respiratory status with 95% confidence intervals

Table 8.3 and figure 8.2 show the variation of FEV1 according to respiratory disease status.

Subgroup analysis of CV_T was also performed by sex, age, ethnicity, geographical region, Townsend deprivation score, BMI, presence of comorbidities, alcohol consumption and smoking status (see Appendix 8.1 figs 18.1 to 18.18).

For FEV₁, higher CV_{TS} were seen in the following subgroups: males compared to females; ex-drinkers and lifelong teetotallers compared to current drinkers; current smokers and ex-smokers compared to participants who have never smoked; participants of white or missing ethnicity compared to other ethnicities; participants with higher deprivations scores (although number of participants in some of these subgroups was low and the error bars were large). There was an increase in CV_T with increasing age. Participants with higher BMIs tended to have lower CV_{TS} . For subgroups based on comorbidities, COPD and heart failure had the highest CV_{TS} and no comorbidity or diabetes had the lowest CV_{TS} . There was no obvious pattern in the differences of CV_T seen between patients from different regions.

Similar findings were seen for FVC except: there was little difference in CV_T between sexes; black ethnicity had a higher CV_T compared to other ethnicities, (but this finding is again hampered by low patient numbers in some ethnic subgroups); a less clear increase in CV_T with deprivation score; no obvious difference in CV_T with BMI.

Figures 8.3 and 8.4 show heat maps of coefficient of variation for FEV₁ and FVC respectively when analysed by multiple subgroups. Figures 18.19 and 18.20 show the same heat maps using long-term data, i.e. not confined to six months from the first measurement. All four heat maps show that respiratory diagnosis has the largest effect on variation.

			Sex	M	M	F	F
			Age	20-60	>60	20-60	>60
Respiratory disease status	Smoker status	BMI					
No respiratory disease	NS	<30	0.152	0.153	0.178	0.142	
		>30	x	x	x	0.155	
	S	<30	0.141	0.166	0.108	0.121	
		>30	0.173	0.193	0.134	0.115	
Asthma	NS	<30	0.160	0.209	0.154	0.195	
		>30	x	0.204	0.137	0.192	
	S	<30	0.235	0.255	0.223	0.249	
		>30	0.243	0.207	0.195	0.229	
COPD	NS	<30	x	0.313	x	0.292	
		>30	x	0.317	x	0.217	
	S	<30	0.327	0.305	0.290	0.277	
		>30	0.258	0.252	0.231	0.244	

Figure 8.3 Heat map showing the coefficient of variation of FEV1 expressed as a decimal fraction for multivariable analysis according to sex, age, smoking status and BMI.

Age is in years. NS=non-smoker, S=current or former-smoker. Groups with < 10 participants were excluded and are marked with an x.

			Sex	M		F	
			Age	20-60	>60	20-60	>60
Respiratory disease status	Smoker status	BMI					
No respiratory disease	NS	<30	0.100	0.155	0.191	0.114	
		>30	x	x	x	0.138	
	S	<30	0.079	0.125	0.124	0.259	
		>30	x	0.128	0.149	0.111	
Asthma	NS	<30	0.115	0.148	0.129	0.181	
		>30	0.102	0.164	0.090	0.114	
	S	<30	0.136	0.164	0.129	0.166	
		>30	0.182	0.148	0.177	0.153	
COPD	NS	<30	x	0.196	x	0.196	
		>30	x	0.176	x	0.125	
	S	<30	0.179	0.171	0.165	0.168	
		>30	x	0.170	x	0.161	

Figure 8.4 Heat map showing the coefficient of variation of FVC expressed as a decimal fraction for multivariable analysis according to sex, age, smoking status and BMI.

Age is in years. NS=non-smoker, S=current or ex-smoker. Groups with < 10 participants were excluded and are marked with an x.

Variation increased with decreasing mean and median FEV1 and FVC levels
i.e. variation increased with worsening disease status, see figs 8.5 and 8.6.

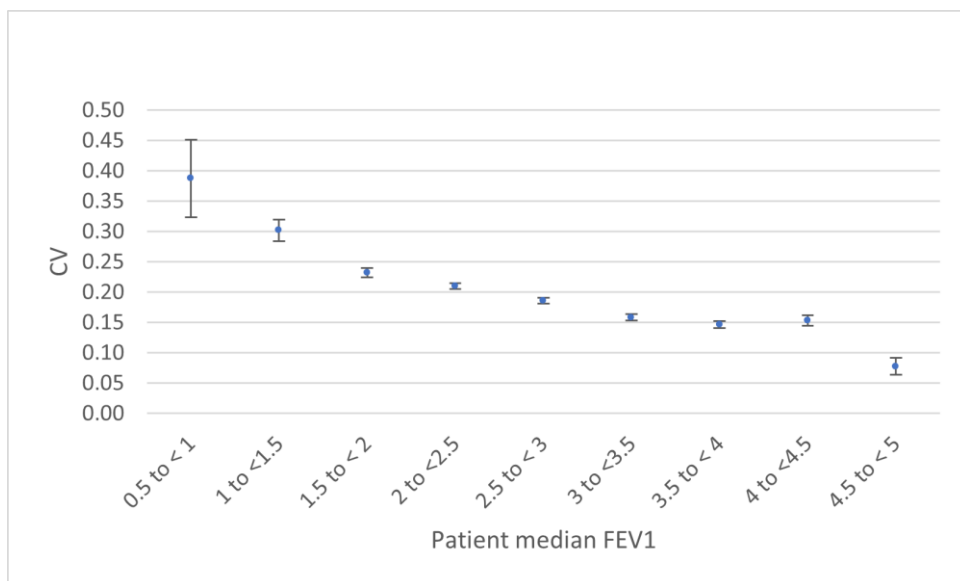


Figure 8.5 Coefficient of variation of FEV1 by patient median with 95% confidence intervals

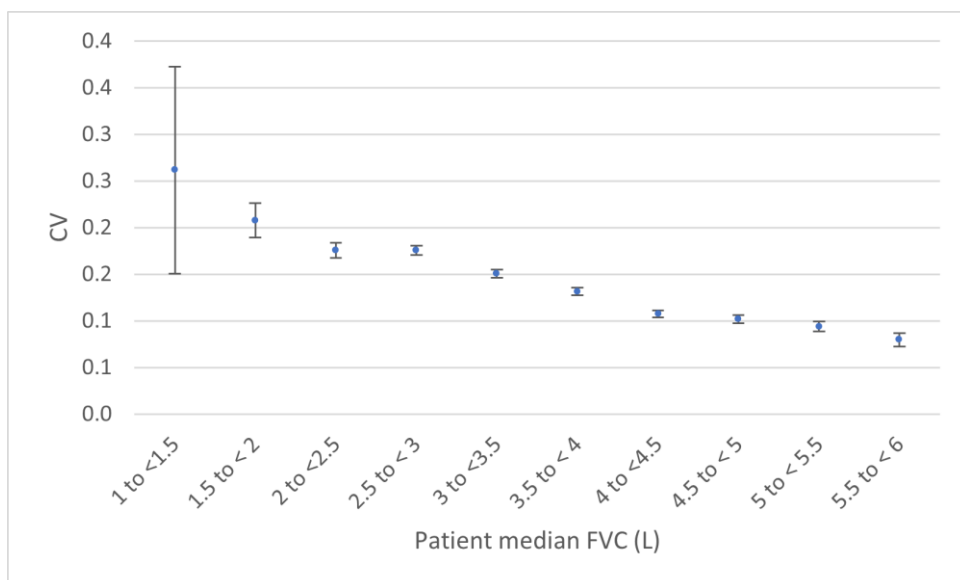


Figure 8.6 Coefficient of variation of FVC by patient median with 95% confidence intervals

first result		0 to < 0.5	0.5 to < 1	1 to < 1.5	1.5 to < 2	2 to < 2.5	2.5 to < 3	3 to < 3.5	3.5 to < 4	4 to < 4.5	4.5 to < 5	5 to < 5.5		Difference -0.5 or less	Difference +.5 or more
0 to < 0.5		30.45	45.10	14.46	5.23	2.28	1.14	0.76	0.38	0.19	0.00	0.00			69.55
0.5 to < 1		3.19	63.01	24.77	5.12	2.19	1.05	0.54	0.08	0.02	0.01	0.00		3.19	33.78
1 to < 1.5		0.37	14.10	60.35	18.79	3.87	1.55	0.77	0.15	0.03	0.00	0.01		14.47	25.17
1.5 to < 2		0.21	2.68	22.16	55.13	14.60	3.35	1.42	0.34	0.09	0.01	0.01		25.05	19.81
2 to < 2.5		0.12	1.65	5.97	25.44	49.75	13.07	2.99	0.80	0.18	0.03	0.00		33.18	17.07
2.5 to < 3		0.12	1.20	3.71	9.26	27.76	44.38	10.68	2.31	0.47	0.10	0.01		42.04	13.57
3 to < 3.5		0.09	1.17	3.65	6.75	10.96	27.86	37.54	9.84	1.74	0.28	0.09		50.47	11.94
3.5 to < 4		0.09	0.90	2.31	4.87	8.37	11.19	27.25	35.97	7.65	1.15	0.21		54.98	9.01
4 to < 4.5		0.26	0.52	1.55	2.07	4.92	9.44	13.58	23.93	35.32	6.99	1.03		56.27	8.02
4.5 to < 5		0.00	2.39	0.48	2.39	1.91	7.66	10.05	10.05	22.97	33.01	8.61		57.89	8.61
5 to < 5.5		0.00	1.56	3.13	1.56	4.69	10.94	6.25	14.06	14.06	15.63	25.00		71.88	0.00

Figure 8.7 Probability of second FEV1 result being 0.5l or more different than first result.

The last two columns show the probability of a second result being at least 0.5L less or more than the previous result.

first result		0 to < 0.5	0.5 to < 1	1 to < 1.5	1.5 to < 2	2 to < 2.5	2.5 to < 3	3 to < 3.5	3.5 to < 4	4 to < 4.5	4.5 to < 5	5 to < 5.5		Difference -0.5 or less	Difference +.5 or more
0 to < 0.5		13.92	16.46	15.19	10.13	17.72	12.66	5.06	3.80	1.27	2.53	1.27			86.08
0.5 to < 1		1.74	27.17	39.70	17.62	8.19	4.09	0.99	0.25	0.25	0.00	0.00		1.74	71.09
1 to < 1.5		0.28	6.16	41.18	34.55	10.88	4.73	1.39	0.68	0.15	0.00	0.00		6.44	52.38
1.5 to < 2		0.12	1.31	13.95	46.48	26.70	7.79	2.19	0.98	0.34	0.12	0.03		15.38	38.13
2 to < 2.5		0.07	0.38	3.51	19.81	45.84	21.96	5.52	1.98	0.71	0.19	0.01		23.77	30.37
2.5 to < 3		0.04	0.19	1.22	5.58	24.09	43.04	18.45	5.26	1.53	0.41	0.11		31.12	25.76
3 to < 3.5		0.03	0.08	0.46	2.35	8.12	25.75	39.19	17.80	4.68	1.17	0.32		36.80	23.97
3.5 to < 4		0.02	0.07	0.42	1.44	3.69	10.36	26.97	36.76	15.67	3.62	0.73		42.97	20.02
4 to < 4.5		0.10	0.04	0.23	0.56	2.10	5.56	11.64	26.09	36.22	14.06	2.68		46.32	16.74
4.5 to < 5		0.00	0.00	0.20	0.80	0.88	2.80	5.64	12.63	29.38	33.77	10.95		52.32	10.95
5 to < 5.5		0.10	0.00	0.10	0.39	0.49	1.96	2.46	5.99	13.46	28.09	33.30		53.05	0.00

Figure 8.8 Probability of second FVC result being 0.5l or more different than first result.

The last two columns show the probability of a second result being at least 0.5L less or more than the previous result.

Figs 8.7 and 8.8 show a heat map with data on the probability of the result of a second test given the result of an initial test.

8.5.1.2 *Sensitivity analyses*

Sensitivity analyses performed for CV_T by year, quarter, number of measurements, implausible values, median FEV1 and FVC level, number of days between measurements, results coded as L v % and a crude estimate of CV_T using an arithmetic method rather than a linear regression model for calculation, and using long-term measurements are reported in Appendix 8.2 figs 18.21 to 18.37.

For FEV1 there was little difference in CV_T when calculated by a simple arithmetic method compared to the linear regression model. CV_T was lower when calculated for results recorded in units of percent predicted compared to litres, but the number of participants for percent predicted were lower (4,412 for L and 515 for % predicted). There was no evidence of marked seasonality or a trend over years for the population CV_T . Data stratified by patient median or mean was similar except for very low values. A smaller interval between measurements was associated with a higher CV_T , and a slight increase in CV_T with a larger number of measurements, consistent with more measurements being performed when disease is less stable.

Although the main analyses were performed on measurements within six months of each other in order to remove the possibility of decline in lung function affecting the results, there was little difference in findings from the short-term results when a long-term analysis involving all recorded measurements was performed (See tables 18.5, 18.6 and figs 18.18, 18.19 and 18.37).

FVC CV_T had similar findings to the analyses for FEV1.

8.6 Discussion

8.6.1 Key results

This is one of the largest studies of between-day within-individual variation of FEV1 to date, with 4,412 participants for FEV1 and 3,567 for FVC. A previous systematic review found 104 studies reporting within-individual variation of FEV1 and/or FVC, but most had small numbers of participants and/or small numbers of repeat measurements (chapter four).

The previous largest study of between-day within-individual variation of FEV1 and FVC was performed in 1995, and consisted of only two measurements [250]. The systematic review in this thesis reported median CV_{TS} of 0.028 for FEV1 and 0.029 for FVC for healthy populations; 0.068 for FEV1 and 0.068 for FVC for patients with COPD; and 0.030 for FEV1 and 0.024 for FVC for patients with asthma.

In the current study, the overall coefficient of variation for FEV1 was 0.224 (95% CI 0.221 to 0.228) with a mean of 2.03L and for FVC was 0.152 (95% CI 0.150 to 0.155) with a mean of 3.03L. This is much higher than seen in the systematic review (see chapter four). For participants with COPD, asthma or both conditions, the CV_{TS} of FEV1 and FVC were also much higher than seen in the systematic review.

In this study, higher CV_T of spirometry measures was associated with having a higher deprivation score, being a current or ex-smoker, being older and having COPD or heart failure. Factors such as age, presence of COPD and smoking history are likely to affect the pathological variation in spirometry measures due to their effect on lung function and may also to some extent be interdependent (e.g.

deprivation and smoking). Participants with higher BMI tended to have lower variation. Previous studies are contradictory as to whether BMI has an effect on spirometry results, possibly related to whether populations were healthy or had COPD. However, one study showed that median FEV1 was higher in patients with higher BMI. Since in this study, a higher median FEV1 is associated with a lower variation, the lower variation in patients with higher BMI may relate purely to the difference in median FEV1 [359, 360]. The median age for participants in this study was relatively high at 72 (for the FEV1 analysis) which might be associated with an increased pathological variability due to increased incidence and severity of COPD and other comorbid conditions in older patients.

CV_T was seen to increase as mean FEV1 or FVC decreased. This suggests that patients with more severe disease status had increased variation of spirometry values. This is an important finding, since patients with more severe disease may have less accurate results, making treatment decisions harder.

The heat maps (figs 8.3 and 8.4) show that disease status is an important factor associated with CV_T , with patients with COPD having higher CV_T than healthy patients, and patients with asthma being intermediate.

8.6.2 Strengths and limitations

One strength of this study is the large size of the dataset, although the number of participants was greatly reduced by the decision to restrict analyses to short-term measurements. However, sensitivity analyses with longer term measurements involving much higher numbers of participants gave similar results (see tables 18.4 and 18.5). It is also relevant that the data is real-world data taken from the clinical records of primary care practices in the UK, and not measurements taken under

ideal or experimental conditions. The Biological Variation Data Critical Appraisal Checklist (BIVAC) recommends biological variation studies are performed on patients in steady state, with “preanalytical procedures described and standardized to minimize preanalytical variation” [154]. This is important when using biological variation data to set analytical performance specifications, but is less relevant to clinical decision-making. Thus the real-world nature of the data in this study is both a strength and a weakness. Another weakness with the data is that it is hard to account for confounding by indication, for example if less well-controlled patients tend to have more frequent measurements. Nevertheless, the large disparity between this study and previously published results and the large number of participants compensate for this problem.

Inclusion of the entire dataset meant the only potential bias introduced into the results (that was not already present in the database) could come from data cleaning (e.g. exclusion of implausible values) and the specification that at least four measurements must be included. However, this latter specification, although aimed at increasing the accuracy of the variation calculation, could introduce confounding by indication since patients with more frequent tests might have more severe/less stable disease and hence a higher degree of pathological variation. Furthermore, it has been shown that the ordering of any test by a clinician increases the risk of the disease which the test is investigating, since the presence of a test in a patient record suggests there was a suspicion that patient was at risk of the disease. This could cause some degree of selection bias when compared to results obtained from healthy populations in controlled biological variation studies [338].

Assumptions were made for the repeat results calculation, namely that the change between first and second result is representative of changes in the result in subsequent calculations, which may not be the case.

Another weakness with the data is error associated with data entry and recording into the clinical databases. Data cleaning prior to analysis partially compensates for this. For example implausible values were removed, taking the cut-off points from published data (e.g. [331]) and analysis of the IMRD data itself, duplicate results on the same day were removed, participants with less than four repeated measurements were eliminated, and patients with four or more identical results were removed, which is a highly unlikely finding. However, it should be noted that if some patients genuinely had four identical results, removing these results from the analysis would artefactually increase the estimated variation.

8.6.3 Interpretation and generalisability

This study shows that the variation of FEV1 and FVC in this population of primary care patients in the UK is much higher than results for variation previously reported in the literature. This could be because the participants might be less likely to be stable, (although good practice suggests spirometry should only be performed in stable patients), and because the pre-analytical testing conditions were not as rigorously controlled as would be the case in a clinical study. Further, the expertise of those performing the testing may be greater in a clinical trial than in general practice, or the technicians may have greater quality control oversight [241]. This means that the results of this study are likely to be more generalisable to a primary care population.

The findings of this study have clinical implications for diagnosis and monitoring of respiratory diseases such as COPD.

For diagnosis, not taking within-individual variation into account could lead to false positives and the use of inappropriate interventions such as glucocorticoids which could lead to adverse effects, while false negatives could lead to appropriate treatment being withheld, delaying diagnosis and causing increased patient morbidity. Similarly, for monitoring, a measured FEV1 or FVC much lower or higher than the real value might lead to inappropriate treatment or withholding of treatment, leading to suboptimal outcomes. This problem could be accentuated when referring to changes from previous results, especially if the variation at the time of measurement is in the opposite direction to the previous measurement (e.g. lower than the real value by chance on the first measurement and higher than the real value by chance on the second measurement).

The concept of minimal clinically important difference (MCID) in spirometry is highly complicated, since MCID differs with diagnosis, severity of disease and what is considered clinically important (for example severity of symptoms versus long-term outcomes) [361]. Major trials have used values for MCID ranging from 45 to 180 mL. However, a pre-bronchodilator FEV1 improvement of around 100 mL is generally considered to correlate with other important clinical outcomes [362].

As figure 8.7 shows, based on estimates from the dataset, the probability of a patient with an initial FEV1 measurement of 1.0 to 1.5L having a second FEV1 level of 0.5 L or less is 0.251, suggesting around one in four patients with an initial FEV1 will by chance have a second measurement more than five times the MCID.

The CV_T of this study can be used to calculate the probability with which a measured value reflects the true mean of the patient, (assuming normal distribution and using a Z-score of difference between means/standard deviation). If we take a hypothetical patient with a true FEV1 of 2.75L the CV_T from the data in this study is approximately 0.182. This suggests a probability of 0.32 that the measured result will be 0.5L higher or lower than the true mean, which is five times the minimal clinically important difference. It has been shown that spirometry results influence decisions regarding medication in the management of COPD and asthma in primary care [363]. It can therefore be seen that the high variation of between-day spirometry measurements could lead to incorrect treatment decisions being made.

Guidelines for COPD and asthma diagnosis are generally based on prediction equations given as per cent predicted rather than absolute values, but few participants had their FEV1 and FVC recorded as per cent predicted in this study, making it hard to make specific estimates for how the variation found in this study would affect clinical outcomes. However, FEV1 is used to make clinical judgements regarding individuals as well as regulatory and guideline recommendations [361], and it is clear that the high level of between-day within-individual variation seen here would make measurements taken on a single day potentially unrepresentative of a patient's true status.

Clinicians should therefore be aware of the possibility for misdiagnosis and inappropriate management of COPD when relying on single or small numbers of between-day spirometry measures, and authors of clinical guidelines should take variation into account when making diagnosis, monitoring and treatment recommendations.

Note that some of the underlying diseases for which tests are performed have a higher variability than others, that is pathological variation will comprise a higher proportion of the total variability. For example, asthma is considered a highly variable disease, while well-controlled diabetes may have much lower variability.

Further research that addresses confounding by a change in frequency or timing of testing with increased disease severity or variation would be useful. Studies into how variation in this population affects clinical outcomes such as mortality and whether reducing variation with improved testing and management protocols improves outcomes is also important.

8.7 Conclusions

This study suggests that the real-world within-individual total measured variation of the spirometry measures FEV1 and FVC are much higher than previously reported in the literature. It also demonstrates that within-individual variation increases with decreasing FEV1 and FVC, i.e. with increasing disease severity, consistent with the findings for HbA1c and CRP.

The next chapter uses the CPRD database to estimate the variation of depression symptom measurement instruments, since these scores are not recorded in the IMRD database. Given how little data on this subject was found in the systematic review, this will be the first meaningful estimate of within-individual measured variation of depression measurement instruments.

9 Chapter nine: Within-individual variation of depression symptom measurement instruments in primary care: a retrospective cohort study

9.1 Abstract

Background

Depression symptom measurement instruments such as PHQ-9, HADS and BDI are used to aid in the diagnosis and management of depression. Variation in within-individual depression symptom measurements may affect their clinical utility but estimates of this variation are currently almost non-existent.

Methods

Data on PHQ-9, HADS and BDI results and sociodemographic, lifestyle and comorbidity covariates were extracted from the CPRD database using the DEXTER tool. A minimum of four measurements in the same individual was the only inclusion criterion. Within-individual measured variation was calculated as a coefficient of variation (CV) using a linear regression random effects model for the whole population and various subgroups.

Results

195,171 participants had at least four measures of PHQ-9 after impossible values were excluded. For HADS, this figure was 8,231 participants and for BDI 215, making this by far the largest study of variation of depression symptom

measurement instruments to date. The overall CV_T (95% confidence interval) was 0.412 (0.411 to 0.413) for PHQ-9, 0.370 (0.365 to 0.374) for HADS and 0.763 (0.701 to 0.826) for BDI. In contrast to studies of variation of other tests, variation decreased with increased mean patient score (i.e. variation decreased with worsening disease severity). This is likely a mathematical function of the integer scale used in these tests.

Strengths and limitations

The large number of participants and the real-world nature of the results are important strengths of this study. Weaknesses included difficulties accounting for confounding by indication.

Conclusions

Estimated within-individual variation in this analysis of real-world data is very high and has not previously been reported. Variation decreases with patient mean score i.e. with more severe disease status which is a counter-intuitive finding and is likely related to the measurement scale. The high within-individual variation of depression measurement instruments and the higher estimated variation in less severely affected patients have important implications for the diagnosis, monitoring and clinical decision-making for depression.

9.2 Introduction

Depression is a serious, recurrent disorder that is one of the most common causes of disability worldwide [364], with around one in five people experiencing major depression at some point in their lifetime [365]. Around one in ten patients presenting to primary care practices have depressive symptoms [366].

There is no single objective clinical test for depression, so the gold standard method for diagnosis is the structured clinical interview, such as the Structured Clinical Interview for the Diagnostic Statistical Manual-5 Clinical Version (SCID-5-CV) which has been shown to have excellent reliability with kappa values largely >0.7 , and high specificity for the disorder [299]. Unfortunately, the SCID-5-CV takes a long time to administer, making it impractical for use in primary care. A number of depression symptom measurement instruments that take the form of Patient Reported Outcome Measures (PROMs) have been devised and validated to assist with screening for depression and assessing severity such as the Patient Health Questionnaire-9 (PHQ-9), the Becks Depression Inventory (BDI) and the Hospital Anxiety and Depressions Score (HADS) (see Chapter 4, Appendix 4.1 for description of these instruments). National Institute for Health and Care Excellence (NICE) guidance recommends the use of a screening instrument such as PHQ-9 for assessing the severity of depression [367]. Historically depression was classified on the basis of PHQ-9 score as 0-4, minimal; 5-9 mild; 10-14 moderate; 15-19 moderately severe; 20-27 severe [368]. However, the most recent NICE guidelines classify depression based on PHQ-9 scores simply as less severe (less than 16) and more severe (16 or more) [367].

The Meaningful Change Threshold, an anchor-based method of estimating a clinically relevant change in disease severity, which is considered more relevant to the patient than the Minimal Clinically Important Difference (MCID) has been estimated as 6 points on the PHQ-9 scale [369]. Another study suggests that a difference in PHQ-9 score is clinically meaningful at 3 points and clinically substantial at 6 points [370]. The MCID for BDI has been estimated as a between 17.5% and 32% change from baseline, depending on initial severity, duration and refractoriness to medication [371, 372]. The MCID for HADS has been estimated as 1.7 in patients with cardiovascular disease [373], 2.7 to 8.5 in patients undergoing total knee or hip arthroplasty [374], 5.7 for patients with rotator cuff disease and shoulder prosthesis [375] and approximately 1.5 in patients with chronic obstructive pulmonary disease [376].

Measurement-based care in depression, which often utilises PROMs, has been shown to lead to improved remission rates [377]. Measurement based care is defined as “the routine measurement of symptoms and side effects at each treatment visit and the use of a treatment manual describing when and how to modify medication doses according to these measurements,” and its use is recommended in American Psychiatric Association Practice Guidelines [378]. Research is ongoing to optimise the effectiveness of measurement-based care in depression [378].

It has been asserted that the use of depression symptom measurement instruments such as the PHQ-9 are underused in practice for the monitoring of depression treatment, and that since measurement-based care has a strong evidence base, increased use of the PHQ-9 for monitoring is likely to improve depression outcomes [379]. However, for PROMs to be truly useful, the results must

be accurate and reliable. While the PHQ-9, BDI and HADS have all been validated for use in multiple languages, settings and disease statuses (e.g. [368, 380]) little attention has been paid to how the results of these tests vary within an individual over time.

Within-individual variation of a biological or clinical measurement is broadly composed of three parts. Pre-analytical variation is variation due to varying factors prior to a measurement. This is well-described for biochemical measurands, with factors such as season, time of day, menstrual cycle and altitude contributing to this component of variation, as well as factors affecting a sample between collection and measurement such as storage and preparation [381]. To my knowledge, influences on the pre-analytical variation of depression screening instruments have not been explored, but it would be plausible that factors such as quality of previous night's sleep, hunger and acute emotional stressors as well as test-related factors such as the demographic and demeanour of the clinician administering the test would affect this component of variation. Analytical variation is the variation introduced by imprecision in the measuring process. Biological variation is the within-individual variation, influenced partly by predictable factors such as season [382], time of day or monthly hormone cycles, but also to a large extent by chance. The sum of these three types of variation gives the total within-individual variation [15]. The higher the within-individual variation, the lower the probability that a single measurement is reflective of the true mean. In this study, it is not possible to separate pre-analytical and analytical variation from biological variation, so within-individual coefficient of variations (CV_T) is taken to refer to total within-individual variation i.e. the sum of analytical, pre-analytical and biological variation.

A database of biological variation exists that lists the variation of commonly measured clinical laboratory tests [326]. However, non-laboratory clinical tests such as PROMs are outside the scope of this database, and to my knowledge, no similar database exists for non-laboratory clinical tests.

A recent systematic review showed there was almost no useful information about the long-term variation of PHQ-9, BDI or HADS. “Although 41 studies with a total of 5,116 participants met the inclusion criteria, most studies had only two measurements per individual, less than two weeks apart and there were no publications at all about within-individual variation in patients diagnosed with depression” ([383] and chapter five). Previous studies and work by the author have shown that variation is higher in patients with a diagnosis related to the measurand or with increasing disease severity, for example in HbA1c ([327] and chapter two) CRP (chapter three) and spirometry (chapter 4). It would therefore be useful to know the within-individual variation of depression symptom measurement instruments and whether there are important differences in variation in different subgroups. Further, it would be helpful to clinicians to attempt to quantify the extent to which variation might lead to a misclassification of diagnosis or severity of depression, or misidentification of a clinically important difference, which would have clinical implications.

9.3 Aims

This study aimed to use real-world data (i.e. data from a clinical source rather than an idealised or experimental study) from The Clinical Practice Research Datalink (CPRD) to estimate the total within-individual variation of the depression

symptom measurement instruments PHQ-9, HADS (depression subscale) and BDI, to compare this to previously published data, to investigate the variation within subgroups of patients and to consider the clinical implications of variation of depression scores in relation to diagnosis, monitoring and management of depression.

9.4 Methods

9.4.1 Study design

A population-based observational retrospective cohort study using anonymised primary care data from the Clinical Practice Research Datalink (CPRD) Aurum database.

The study was reported according to the STrengthening the Reporting of Observational studies in Epidemiology (STROBE) statement for observational studies.

9.4.2 Data sources

Study data was obtained from the Clinical Practice Research Datalink (CPRD) Aurum database. CPRD is a not-for profit service that provides anonymised primary care data for scientific research. The CPRD Aurum database, launched in October 2017, captures routinely collected data from primary care practices in England and Northern Ireland [118]. It has been shown to be broadly representative of the UK general population in terms of socio-economic status and rural-urban classification

[119] and ethnicity [120] and has been shown to have high levels of correctness and moderate to high levels of completeness for selected conditions [121].

The dataset used for this study consisted of 24,176,224 participants registered in 1,491 practices.

9.4.3 Ethical approval

CPRD has obtained generic ethics approval for observational studies using anonymised data from CPRD Aurum from the UK's Health Research Authority Research Ethics Committee (Ref:05/MRE04/87). No additional ethical approval is therefore required for this study. The use of CPRD Aurum data for this study was approved by the CPRD Independent Scientific Advisory Committee on 10th May 2023 (reference no. 23_002708).

9.4.4 Setting

Data was obtained from English and Northern Irish general practices registered with the CPRD Aurum database from inception to 19th July 2023. Data extraction from the database was performed using the Dexter software [122] and was completed on 19th July 2023. The date of the latest entrance was 18th March 2022.

9.4.5 Participants

All participants in the database were eligible for inclusion. Except for sensitivity measures, participants were excluded if they did not have at least four repeat

measurements of PHQ-9, HADS or BDI recorded. Participants were followed until the earliest date of either death, leaving the practice, stopping contributing to the database or the study end date.

9.4.6 Variable choice

The co-primary outcomes were within-individual variation of PHQ-9, HADS and BDI, calculated as the within-individual measured coefficient of variation (CV_T). Other variables included in analyses are sociodemographic characteristics (age, sex, ethnicity, deprivation quintile); lifestyle related characteristics (BMI, smoking, alcohol use); use of antidepressants; comorbidities (table 9.1). The date of each depression measurement symptom scoring and the date of diagnosis where applicable were also extracted. Variables for subgroup analysis were selected as those hypothesised most likely to have an effect on variation or that enabled calculation of variation (e.g. date of measurement). Variables for sensitivity analysis were chosen to assess bias due to factors which might affect the robustness of the results such as statistical methods, number of repeat measurements and unit of measurement.

Code lists for the different variables are available at:

<https://github.com/THINKINGGroup/phenotypes>.

Table 9.1 Variables extracted from database.

Variables extracted	
Test characteristics	Depression score result (PHQ-9, HADS, BDI)
	Date of measurement
Patient characteristics	
Sociodemographic	Age
	Sex
	Ethnicity
	BMI
	Geographical region
Lifestyle factors	Smoker status
	Alcohol consumption status

Diagnoses, medications and comorbidities	Anti-depressant medication use
	Diabetes
	Hypertension
	Hyperthyroidism
	Hypothyroidism
	Ischaemic heart disease
	Heart failure
	Ischaemic stroke
	Haemorrhagic stroke
	All cancers

9.4.7 Bias and Study size

The study size included all participants with sufficient repeat measurements of PHQ-9, HADS or BDI, to maximise sample sizes for the whole population, subgroup analysis and sensitivity analyses and help to compensate for weaknesses inherent in routinely collected medical data. Selection bias was also minimised as no participants were excluded except for those with insufficient repeat measurements.

9.4.8 Data cleaning

Values outside of the ranges of the depression measurement instruments (PHQ-9 0-27; HADS depression subscale 0-21; BDI 0-63) [384] were excluded as being impossible.

9.4.9 Statistical methods

Statistical methods were as reported in the methods of the HbA1c study, chapter 6. Minimal clinically important difference (MCID) was considered to be 6 points on the PHQ-9 scale [370]. The chance of getting a difference of at least this amount between the first and second results were calculated as a percentage by using the dataset to evaluate the number of participants in which every combination of first and

second results was achieved (e.g. 1st=1, 2nd=5; 1st=2, 2nd=7 etc). The probability of each combination of first and second results was then calculated by dividing the number of participants with that combination by the total number of participants who had the first result. For example, if 2,944 participants had a first PHQ-9 result of 4, and 126 of these had a second result of 7, the probability of achieving a second result of 7, given a first result of 4, is $126/2944=0.043$ (4.3%). These results were then used to calculate the probability of a second result being more than the MCID (e.g. 6 for PHQ-9).

Subgroup and sensitivity analyses were performed as for the HbA1c study in chapter 5. Age and sex were complete in the database with no missing data. Participants with missing depression score results were not eligible for inclusion. Where the data for other comorbidities was missing, the comorbidity was assumed not to be present. Participants for whom other data was missing (such as ethnicity or BMI) were excluded from that subgroup analysis or were calculated as a separate group (e.g. patients with missing BMI were excluded from the BMI subgroup analysis but not the ethnicity analysis).

Statistical code for Stata can be found is available from the author on request.

9.5 Results

Participants

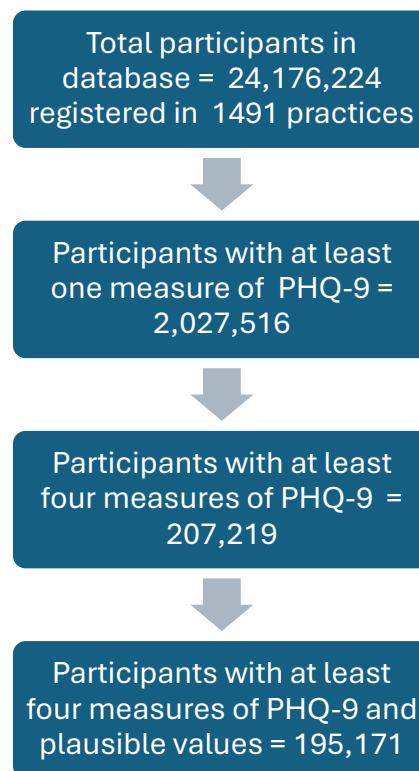


Figure 9.1 Participant selection algorithm for PHQ-9

24,176,224 participants registered in 1,491 practices were available in the database for analysis. Figure 9.1 shows the flowchart of participant selection for PHQ-9. 195,171 participants had at least four measures of PHQ-9 with values 0 to 27; 8,231 participants had at least four measures of HADS with values of 0 to 21; 215 participants had at least four measures of BDI with values of 0 to 63. The median number of measurements per individual was 5 (IQR 4-7) for PHQ-9, 6 (IQR 5-8) for HADS and 5 (IQR 4-7) for BDI.

9.5.1 Descriptive data and outcome data

The overall within-individual coefficient of variation (for all eligible participants) for PHQ-9 was 0.412 (95% CI 0.411 to 0.413) with a median PHQ-9 score of 13. The overall coefficient of variation for HADS was 0.370 (95% CI 0.365 to 0.374) with a median score of 10. The overall coefficient of variation for BDI was 0.763 (95% CI 0.701 to 0.826) with a median score of 18 (fig 9.2).

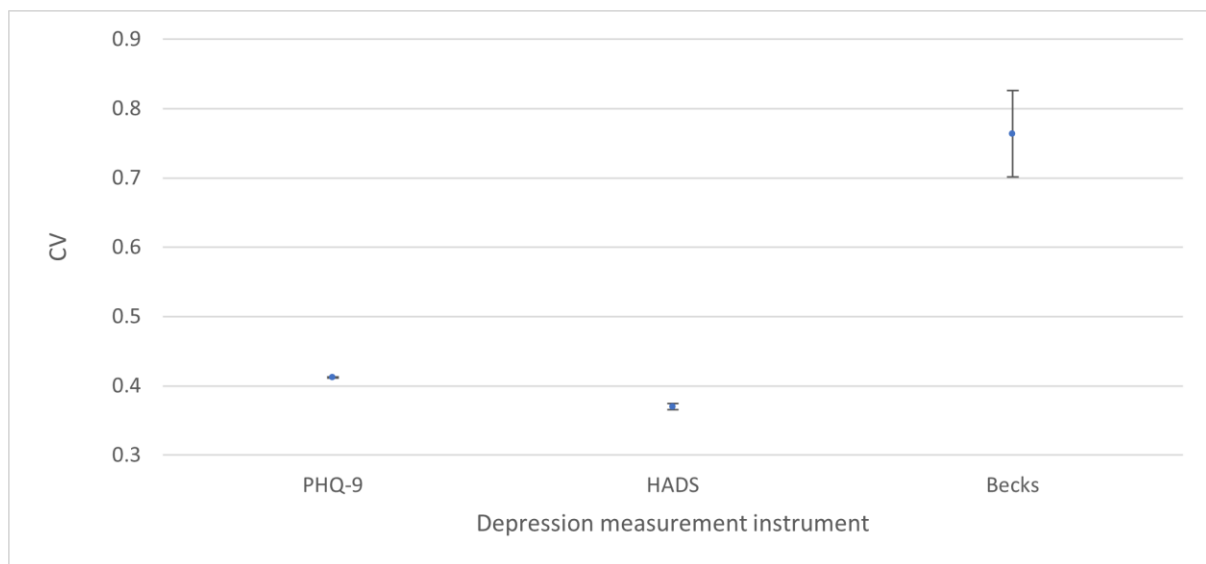


Figure 9.2 Overall CV of PHQ-9, HADS and BDI with 95% confidence intervals.

CV decreased with increasing individual mean and median PHQ-9 scores across the whole study timeframe, (see fig 9.3 and fig 19.13) i.e. variation decreased with worsening disease symptom score. A similar pattern was seen for HADS and BDI CVs stratified by mean, see figs 19.14 and 19.15.

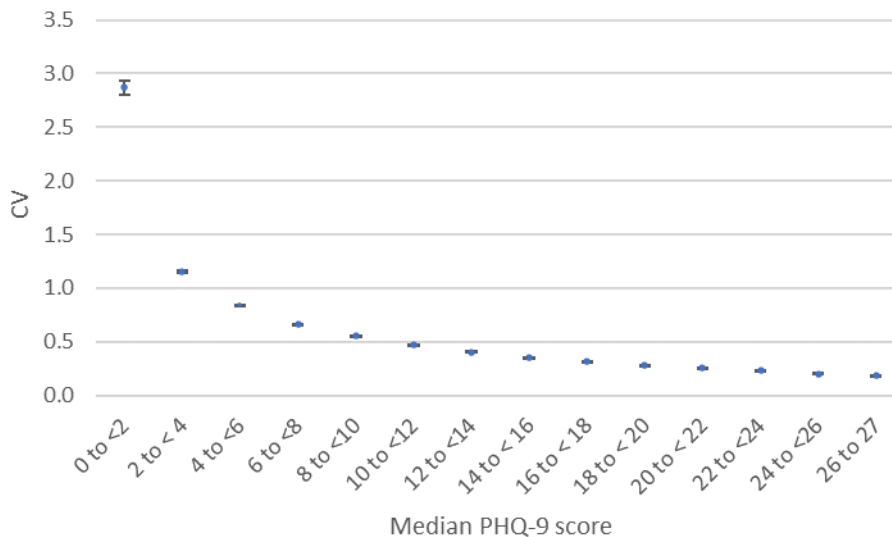


Figure 9.3 Coefficient of variation of PHQ-9 by patient median with 95% confidence intervals.

Table 9.2 shows the participant characteristics for patients with at least four measurements. 195,171 participants were included in the study, with a median age at first measurement of 48 (IQR 36 to 41), 35.3% being male. Data was recorded as “Missing” for ethnicity in 17,209/195,171 (8.8%) participants. Drug data was recorded for 160,739 participants. The median number of years between first and last measurement in the database was 3 (IQR 1 to 6).

9.5.1.1 Subgroup analyses

Table 9.2, table 19.1 and figures 19.1 to 19.8 show the subgroup analyses. For PHQ-9, higher CV_T s were seen in the following subgroups: participants of Black and South Asian ethnicity compared to other ethnicities; participants with BMIs of 25-30 compared to participants with other BMIs; each of the comorbidities analysed compared to the overall CV_T . Patients without comorbidities had a lower CV_T than the overall CV_T . Increasing age was associated with an increasing CV_T . Little

difference was seen between male and female participants, participants from different regions, and participants with different smoking statuses. A longer median time between measurements and a lower PHQ-9 score were associated with a higher CV_T.

Table 9.2 Sociodemographic data and CV of subgroups

TCAD = tricyclic anti-depressant, SSRI = serotonin reuptake inhibitor; MAOi = monoamine oxidase inhibitor. COPD = chronic obstructive pulmonary disease, CKD=chronic kidney disease, T1DM = type 1 diabetes. T2DM= type 2 diabetes

		Mean PHQ-9 score	CV (95% CI)	N (%)
Whole population		12.7	0.412 (0.411 to 0.413)	195,171 (100)
Ethnicity	Black	10.7	0.457 (0.449 to 0.466)	6,744 (3.5)
	Mixed	13.6	0.387 (0.377 to 0.397)	1,538 (0.8)
	Other	13.5	0.388 (0.380 to 0.396)	2,581 (1.3)
	South Asian	9.8	0.493 (0.486 to 0.500)	12,291 (6.3)
	White	12.9	0.408 (0.407 to 0.409)	154,808 (79.3)
	Missing	13.4	0.393 (0.390 to 0.396)	17,209 (8.8)
BMI	<=18.5	13.8	0.376 (0.371 to 0.381)	6,226 (3.2)
	>18.5 to 25	12.9	0.409 (0.408 to 0.411)	58,083 (29.8)
	>25 to 30	12.2	0.429 (0.427 to 0.431)	42,131 (21.6)
	>30 to 35	12.5	0.422 (0.419 to 0.425)	21,244 (10.9)
	>35	12.8	0.405 (0.403 to 0.407)	67,487 (34.6)
Sex	Male	11.9	0.424 (0.422 to 0.426)	68,895 (35.3)
	Female	13.1	0.406 (0.405 to 0.407)	126,268 (64.7)
Age (years)	<=10			0 (0)
	>10 to 20	14.2	0.333 (0.326 to 0.340)	1,951 (1.0)
	>20 to 30	14.5	0.358 (0.356 to 0.360)	24,094 (12.3)
	>30 to 40	14.1	0.380 (0.379 to 0.382)	40,945 (21.0)
	>40 to 50	13.6	0.396 (0.394 to 0.398)	39,770 (20.4)
	>50 to 60	13.4	0.404 (0.402 to 0.406)	38,608 (19.8)
	>60 to 70	11.8	0.447 (0.443 to 0.450)	24,405 (12.5)
	>70 to 80	8.1	0.590 (0.582 to 0.598)	13,804 (7.1)
	>80 to 90	5.9	0.688 (0.673 to 0.702)	8,649 (4.4)
	90 to 100	5.2	0.731 (0.702 to 0.761)	2,794 (1.4)
Median PHQ-9 depression score	0 to <2	1.1	2.869 (2.803 to 2.935)	16491 (8.4)
	2 to < 4	4.0	1.152 (1.136 to 1.168)	6241 (3.2)
	4 to <6	6.1	0.837 (0.828 to 0.846)	7766 (4.0)
	6 to <8	7.8	0.662 (0.657 to 0.667)	10777 (5.5)
	8 to <10	9.4	0.554 (0.551 to 0.557)	15145 (7.8)
	10 to <12	11.0	0.471 (0.468 to 0.473)	19606 (10.0)
	12 to <14	12.6	0.406 (0.404 to 0.408)	23115 (11.8)
	14 to < 16	14.3	0.357 (0.356 to 0.359)	24651 (12.6)
	16 to < 18	15.9	0.319 (0.318 to 0.320)	23125 (11.8)
	18 to < 20	17.6	0.286 (0.284 to 0.287)	20061 (10.3)

	20 to < 22	19.4	0.253 (0.252 to 0.255)	14876 (7.6)
	22 to <24	21.2	0.228 (0.226 to 0.229)	8521 (4.4)
	24 to <26	22.9	0.203 (0.201 to 0.205)	3869 (2.0)
	26 to 27	24.6	0.180 (0.176 to 0.184)	927 (0.5)
Median days between measurements	0 to < 4	12.2	0.360 (0.338 to 0.382)	361 (0.2)
	4 to < 7	14.1	0.362 (0.340 to 0.384)	271 (0.1)
	7 to < 14	12.8	0.370 (0.365 to 0.376)	3,992 (2.0)
	14 to < 21	13.1	0.368 (0.364 to 0.371)	8,387 (4.3)
	21 to < 28	13.5	0.369 (0.366 to 0.373)	9,210 (4.7)
	28 to < 60	13.6	0.383 (0.381 to 0.385)	39,980 (20.5)
	60 to < 90	13.7	0.398 (0.396 to 0.401)	21,662 (11.1)
	90 to < 120	13.3	0.410 (0.407 to 0.414)	14,080 (7.2)
	120 to < 150	13.5	0.410 (0.406 to 0.414)	10,240 (5.2)
	150 to < 180	13.6	0.406 (0.401 to 0.410)	8,370 (4.3)
	180 to < 270	12.6	0.427 (0.424 to 0.431)	19,785 (10.1)
	270 to < 365	10.8	0.464 (0.459 to 0.470)	16,920 (8.7)
	365 to < 455	8.8	0.525 (0.517 to 0.533)	14,944 (7.7)
	455 to < 545	11.1	0.463 (0.455 to 0.470)	8,013 (4.1)
	545 to < 635	12.1	0.440 (0.432 to 0.448)	5,090 (2.6)
	635 to < 730	13.0	0.422 (0.414 to 0.430)	3,730 (1.9)
	>730	13.6	0.417 (0.412 to 0.422)	10,115 (5.2)
Depression status	Depression diagnosis	13.7	0.391 (0.389 to 0.393)	36,141 (18.5)
	No depression diagnosis	11.7	0.435 (0.433 to 0.436)	116,421 (59.7)
	Missing	-	-	42,609 (21.8)
Anti-depressant use	No drugs	11.5	0.442 (0.440 to 0.444)	104,509 (53.5)
	TCAD	13.8	0.385 (0.382 to 0.388)	13,278 (6.8)
	SSRI	13.9	0.383 (0.381 to 0.385)	34,548 (17.7)
	MAOi	13.9	0.374 (0.346 to 0.402)	179 (0.1)
	Other antidepressant	14.4	0.365 (0.361 to 0.369)	8,279 (4.2)
Comorbidities	No comorbidity, no depression	12.7	0.411 (0.409 to 0.402)	76,975 (39.4)
	Depression, no comorbidity	14.0	0.384 (0.382 to 0.387)	21,637 (11.1)
	At least one comorbidity, no depression	10.2	0.482 (0.478 to 0.486)	36,853 (18.9)
	Depression, at least one comorbidity	11.9	0.430 (0.429 to 0.432)	70,892 (36.3)
	Depression	13.7	0.391 (0.389 to 0.393)	36,582 (18.7)
	Hypertension	8.6	0.549 (0.541 to 0.556)	15,916 (8.2)
	Asthma	13.2	0.397 (0.395 to 0.400)	27,494 (14.1)
	COPD	10.0	0.482 (0.464 to 0.500)	1,707 (0.9)
	Hyperthyroidism	11.8	0.442 (0.427 to 0.457)	1,245 (0.6)
	Hypothyroidism	11.2	0.458 (0.449 to 0.467)	4,216 (2.2)
	CKDstage3 to 5	7.9	0.583 (0.549 to 0.616)	979 (0.5)
	Heart failure	7.6	0.587 (0.547 to 0.627)	760 (0.4)
	Ischaemic heart disease	6.7	0.643 (0.626 to 0.660)	5,643 (2.9)
	Ischaemic stroke	9.1	0.556 (0.510 to 0.602)	388 (0.2)
	Haem. Stroke	10.8	0.489 (0.447 to 0.531)	244 (0.1)
	Cancer	12.1	0.435 (0.430 to 0.440)	11,325 (5.8)
	T1DM or T2DM	6.9	0.657 (0.643 to 0.671)	8,540 (4.4)

Figures 9.4 and 9.5 show that variability appears to be higher with a diagnosis of no depression and with no anti-depressant drug use. This is likely to be associated with lower PHQ-9 scores which appear to independently affect the CV (see discussion).

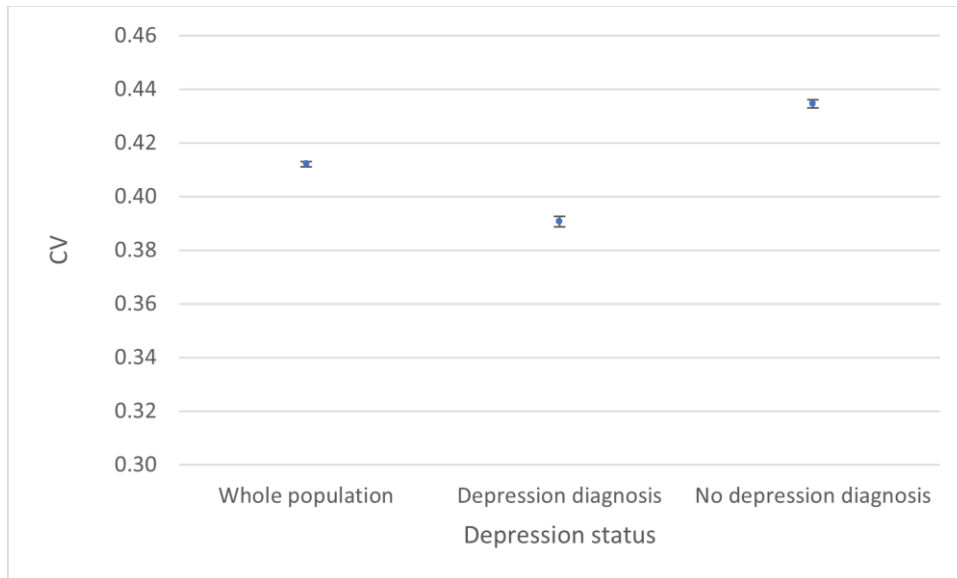


Figure 9.4 Coefficient of variation of PHQ-9 score by depression diagnosis

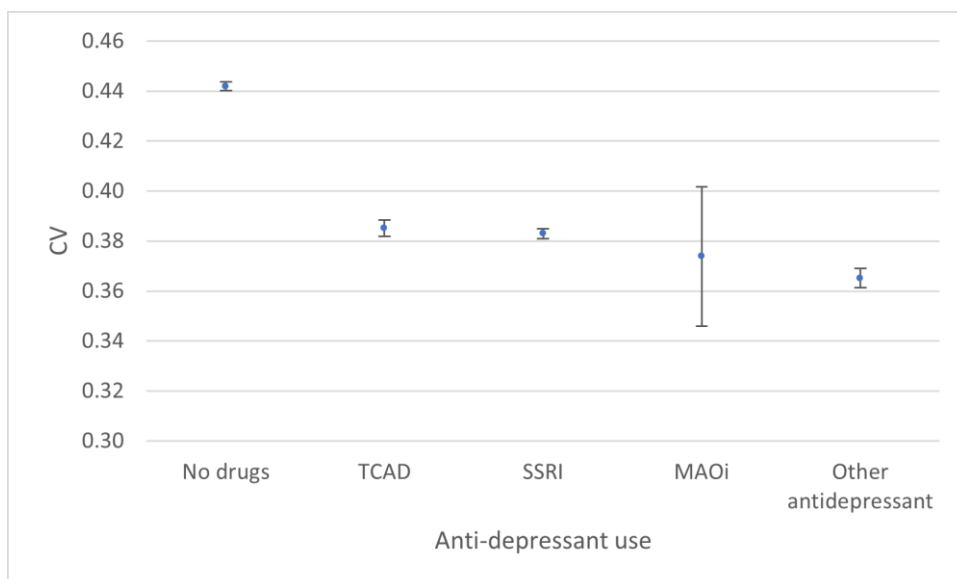


Figure 9.5 Coefficient of variation of PHQ-9 by anti-depressant use.

TCAD = tricyclic anti-depressant, SSRI = serotonin reuptake inhibitor; MAOi = monoamine oxidase inhibitor.

Median depression score, time between measurements, depression diagnosis, presence of comorbidities, ethnicity, and age appear to have important effects on CV, so these were chosen for multi-subgroup analysis. The results of this are presented in heat maps in figures 9.6 and S9.9. Mean depression score seems to have by far the most marked effect on CV.

			One or more comorbidity	Y	Y	N	N
			Depression diagnosis	Y	N	Y	N
Depression score	Age	Ethnicity					
0 to 4	20 to 60 >60	White		1.600	1.827	1.245	1.452
				1.782	2.090	1.654	1.962
	20 to 60 >60	Other		1.856	2.270	1.615	1.935
				1.648	1.749	1.480	1.806
5 to 9	20 to 60 >60	White		0.680	0.683	0.683	0.662
				0.715	0.745	0.717	0.696
	20 to 60 >60	Other		0.747	0.750	0.685	0.681
				0.730	0.770	0.741	0.733
10 to 14	20 to 60 >60	White		0.451	0.448	0.457	0.444
				0.459	0.457	0.455	0.446
	20 to 60 >60	Other		0.466	0.457	0.446	0.440
				0.456	0.461	0.462	0.440
15 to 19	20 to 60 >60	White		0.309	0.298	0.305	0.298
				0.310	0.309	0.311	0.301
	20 to 60 >60	Other		0.308	0.298	0.303	0.294
				0.296	0.306	0.298	0.309
20 to 27	20 to 60 >60	White		0.171	0.173	0.178	0.172
				0.177	0.185	0.175	0.173
	20 to 60 >60	Other		0.184	0.172	0.171	0.170
				0.185	0.185	0.172	0.185

Figure 9.6 Heat map showing the coefficient of variation of PHQ-9 expressed as a decimal fraction for multivariable analysis according to age, ethnicity, comorbidities, depression diagnosis and PHQ-9 score. Age is in years.

first result	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Difference -6 or less	Difference +6 or more	Difference of 6 more or less	
0	0.74	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00			0.13	
1	0.30	0.23	0.09	0.06	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00			0.23	
2	0.22	0.10	0.21	0.08	0.06	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00			0.26	
3	0.16	0.06	0.09	0.19	0.07	0.05	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00			0.28	
4	0.11	0.04	0.06	0.08	0.14	0.06	0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.00			0.33	
5	0.08	0.04	0.05	0.06	0.07	0.13	0.05	0.05	0.04	0.04	0.04	0.03	0.05	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.00	0.00			0.36	
6	0.06	0.03	0.04	0.05	0.06	0.06	0.12	0.05	0.05	0.04	0.05	0.04	0.04	0.03	0.04	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.06		0.35	
7	0.05	0.02	0.04	0.04	0.06	0.05	0.06	0.13	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.03	0.02	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.00	0.01	0.07		0.32	
8	0.03	0.02	0.03	0.04	0.04	0.05	0.06	0.05	0.12	0.04	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.08		0.32	
9	0.03	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.12	0.05	0.05	0.05	0.05	0.05	0.04	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.11		0.29	
10	0.03	0.01	0.02	0.03	0.03	0.03	0.04	0.04	0.06	0.06	0.13	0.05	0.06	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.12		0.25	
11	0.02	0.01	0.02	0.03	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.11	0.06	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.15		0.23	
12	0.02	0.01	0.02	0.03	0.03	0.03	0.04	0.03	0.05	0.05	0.07	0.05	0.14	0.05	0.06	0.05	0.05	0.05	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.17		0.18	
13	0.02	0.01	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.06	0.06	0.07	0.12	0.06	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.19		0.17	
14	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.06	0.05	0.07	0.06	0.13	0.06	0.06	0.04	0.05	0.04	0.04	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.21		0.14	
15	0.02	0.01	0.01	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.04	0.07	0.06	0.06	0.13	0.06	0.05	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.00	0.01	0.24		0.11	
16	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.03	0.05	0.04	0.06	0.05	0.07	0.06	0.12	0.05	0.06	0.05	0.04	0.04	0.03	0.02	0.02	0.01	0.00	0.01	0.26		0.09
17	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.06	0.13	0.06	0.06	0.05	0.04	0.04	0.02	0.02	0.01	0.01	0.01	0.28		0.07	
18	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.04	0.06	0.06	0.07	0.06	0.14	0.05	0.05	0.05	0.03	0.03	0.02	0.01	0.01	0.01	0.31		0.05	
19	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.05	0.06	0.06	0.07	0.13	0.06	0.05	0.04	0.03	0.03	0.02	0.01	0.01	0.33		0.03	
20	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.08	0.06	0.13	0.06	0.05	0.04	0.03	0.01	0.01	0.01	0.35		0.02
21	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.03	0.04	0.05	0.05	0.05	0.07	0.06	0.07	0.14	0.05	0.05	0.04	0.03	0.01	0.01	0.37		0.01	
22	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.13	0.05	0.05	0.03	0.02	0.01	0.39			
23	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.04	0.06	0.06	0.07	0.07	0.07	0.13	0.07	0.04	0.02	0.02	0.40			
24	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.02	0.03	0.03	0.03	0.04	0.04	0.06	0.06	0.06	0.07	0.08	0.07	0.14	0.04	0.03	0.03	0.42			
25	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.03	0.05	0.04	0.06	0.07	0.07	0.07	0.08	0.15	0.04	0.04	0.43			
26	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.02	0.02	0.04	0.03	0.04	0.05	0.06	0.06	0.07	0.08	0.09	0.13	0.07	0.45		
27	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.04	0.03	0.03	0.03	0.05	0.06	0.07	0.06	0.09	0.06	0.20	0.45			

Figure 9.7 Probability of second PHQ-9 result based on first result.

The last two columns show the chance of the second score being 6 or more higher or lower than first score.

Fig 9.7 gives a heat map with data on the probability of the result of a second test given the result of an initial test. Figure 9.8 shows the probability of a second test score being 6 or more points higher or lower than the first score.

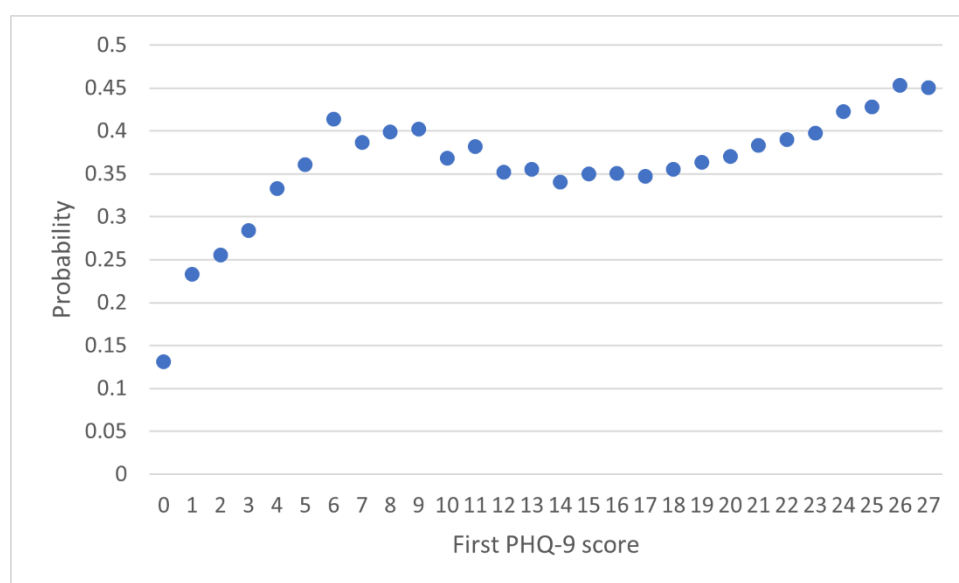


Figure 9.8 Probability of second PHQ-9 score being 6 points or more greater or less than the first score(see methods).

9.5.1.2 Sensitivity analyses

Sensitivity analyses were performed for CV_T by year, quarter, number of measurements, using different thresholds for excluding implausible values, median PHQ-9, log-adjusted mean PHQ-9, number of days between measurements, a crude estimate of CV_T using an arithmetic method rather than a linear regression model for calculation (See Appendix 9.2 figs 19.10 to 19.18).

For PHQ-9 there was little difference in CV_T when calculated by a simple arithmetic method compared to the linear regression model, nor if the data was log-transformed. There was no evidence of marked seasonality or a trend over years for the CV_T . When data was stratified by patient median or mean the results were similar. However, CV_T decreased with number of repeat measurements after

approximately ten measurements. There was little difference in CV_T when comparing number of days between measurements except for a spike at around one year.

9.6 Discussion

9.6.1 Key results

The overall CV_T (95% confidence interval) was 0.412 (0.411 to 0.413) for PHQ-9, 0.370 (0.365 to 0.374) for HADS and 0.763 (0.701 to 0.826) for BDI.

This is by far the largest study of between-day within-individual variation of depression symptom measuring instruments to date, including 129,161 participants with at least four measurements of PHQ-9, 8,231 participants with at least four measurements of HADS and 215 participants with at least four measurements of BDI. This is much larger than the total number of participants in a previous systematic review, which numbered only 5,116, most of which were measured only twice, less than 2 weeks apart. Only one paper expressed the within-individual variation as a CV_T , in which fifteen patients with hypothyroidism symptoms were found to have a CV_T of HADS of 0.416 which is similar this study (chapter five, [306, 383]).

CV_T was seen to decrease as mean and median score increased for all three depression measurement instruments. This is counter-intuitive, and contrary to the findings of how within-individual variation changes with mean or median in other measures such as spirometry (see chapter 8), CRP (see chapter 7) and HbA1c (see chapter 6 and [327]). Generally it might be expected that worsening disease status might be associated with higher variation as, for example, homeostatic mechanisms

work less effectively, or variation in plasma levels of medications fluctuate over time. The fact that this is not the case for the depression symptom measurement instruments is likely to be an intrinsic issue related to the design of the instruments. A fluctuation of one or two points at the lower end of the scale will have a disproportionately higher effect on the CV than a similar fluctuation at the higher end of the scale. This is important since it suggests there may be an intrinsic problem with variation in depression measurement instruments and other PROMs.

A number of differences in variation with different subgroups were found, but it must be considered that confounding by indication may play a role in this, for example participants with worse symptoms may have more frequent measurements. Participants treated with anti-depressant drugs had lower variation than patients not treated, suggesting that anti-depressant drugs reduce depression symptom variation. Higher CV_T of PHQ-9 was associated with being male, being of Black or South Asian ethnicity compared to other ethnicities, being older and having any comorbidity except asthma compared to no comorbidities. These may be genuine independent risk factors for increased variability of depression symptom measurement scores but may be interdependent, for example age and presence of comorbidities.

CV_T of PHQ-9 differed little by year or by quarter, but tended to decrease with increasing number of measurements. There was little difference in CV_T when comparing different measurement intervals except for a spike at around one year. Although there is no obvious explanation for this, it may relate to patients presenting for annual evaluations. There was little difference when the CV_T was calculated by a simple arithmetic method compared to the linear regression method, nor if the data was log-transformed.

However, the coefficient of variation may not fully describe variation in numerical scales such as depression symptom measurement instruments, but there is relatively little work published on how to calculate and describe the variation of integer scales.

Another method to look at the variation of the depression symptom scores is the probability of the value of a second result being within a certain range, given the first result. Fig 9.7 shows that for values more than 6, the probability of a second result being 6 or more points higher than the first result also decreases with mean, supporting the findings of the variation when calculated by CV.

This has an important implication when monitoring depression symptoms, since small fluctuations in more severely depressed patients may be more significant than similar fluctuations in less severely depressed patients. It may also be the case that small changes in depression symptom scores are perceived as being more important to patients with more severe symptoms than mild symptoms.

9.6.2 Strengths and limitations

The major strength of this study is the large size of the dataset, with very little published on this subject to date. Another strength is the fact that the study is based on real-world data taken from the clinical records of UK primary care practices (not just measurements taken under ideal or experimental conditions). It therefore includes all components of variation and is the data which is used to inform clinical decisions.

It should be reiterated that the variation described in this study is the total measured within-individual variation which includes pre-analytical and analytical

variation. However, since clinical decisions are taken based on total variability, it is a strength that this study looks at the total variability (pre-analytical, analytical, biological and pathological) as found in real-world data.

This is the only study of depression symptom measurement instruments that looks at variability within subgroups. However, a weakness with the data is that it is hard to account for confounding by indication, as discussed above. It is also hard to disentangle the effects of multiple risk factors, for example differences in variability by socioeconomic status may reflect differences in mean values. This is also the only study of its kind that considers the clinical implications of the within-individual variability of depression measurement scores. However, it doesn't distinguish between steady-state variation and variation where there has been a material change in the patient's condition. Since depression is not a progressive condition, this should not be a major source of bias. A very short time between measurements was associated with a reduced variability (see fig 19.16). This may indicate a smaller short-term variability, but this may be due to participants remembering their answers from previous tests.

Assumptions were made for the repeat results calculation, namely that the change between first and second result is representative of changes in the result in subsequent calculations, which may not be the case.

Another weakness with the data is error associated with data entry and recording into the clinical databases. Data cleaning prior to analysis partially compensates for this. For example impossible values and duplicate results on the same day were removed.

Inclusion of the entire dataset meant the only potential bias introduced into the results (that was not already present in the database) could come from data cleaning (e.g. exclusion of implausible values) and the specification that at least four measurements must be included. However, this latter specification, although aimed at increasing the accuracy of the variation calculation, could introduce confounding by indication since patients with more frequent tests might have more severe/less stable disease and hence a higher degree of pathological variation. Furthermore, it has been shown that the ordering of any test by a clinician increases the risk of the disease which the test is investigating, since the presence of a test in a patient record suggests there was a suspicion that patient was at risk of the disease. This could cause some degree of selection bias when compared to results obtained from healthy populations in controlled biological variation studies [338].

9.6.3 Interpretation and generalisability

This study is the first to estimate the variation of depression symptom measurement instruments commonly used in general practice. The main conclusion is that the large size of variation found in this study means there is a high probability of clinically important differences being found purely due to chance, especially at the lower end of the measurement scores. Since the finding of increased variation with low scores applies to all three depression symptom measurement instruments investigated in this study (PHQ-9, HADS, BDI), it is likely that this problem is relevant to other PROMs.

The findings of the study have implications for the diagnosis and management of depression since guidelines such as those from NICE recommend the use of a

screening instrument such as PHQ-9 to aid in the diagnosis and monitoring of the condition. Not taking within-individual variation into account during diagnosis could lead to a false positive, with an incorrect diagnosis of depression leading to inappropriate treatment, social stigma, patient distress and effects on insurance. False negatives could lead to missed opportunities for important interventions to reduce patient suffering and reduce the risk of suicidality.

Similarly for monitoring, not taking within-individual variation into account could lead to patients being inappropriately given intensified treatment with possible adverse effects, especially if a previous test result was lower than the real result by chance and the latest test result was higher than the real result by chance. Conversely, false negatives might lead to inadequate interventions, especially if a previous test result was higher than the real result by chance and the latest test result is lower than the real result by chance.

Taking the cut-off point in the latest NICE guidelines of 16 or more to signify severe depression, a patient with a first PHQ-9 score of 16 has a 0.26 probability of their second result being 6 or more points less, suggesting they were no longer severely depressed, and a 0.09 probability of their score being 6 points or more, suggesting a considerable worsening of their depression symptoms. One study concluded that fluoxetine was an effective treatment for major depressive disorder amongst haemodialysis patients after an effect size of -4.7 points on the PHQ-9 [385]. If the 0.26 probability of a second PHQ-9 score being 6 points or more lower than the first is due to chance factors such as regression to the mean, then 26% of patients would have a decrease in PHQ-9 score larger than the effect size of fluoxetine purely due to chance. However, according to one study, most patients assessed with a PHQ-9 score are treated with anti-depressants regardless of result

in UK general practice, so it can't be ruled out that the decrease seen here is actually a treatment effect [386]. Nevertheless, in this study only 35% of patients were treated with anti-depressants at some point during the study period.

An alternative method of assessing the implications of variation of depression scores in practice uses the CV_T of this study to calculate the probability with which a measured value reflects the true mean of the patient, (assuming normal distribution and using a Z-score of difference between means/standard deviation). If we take a hypothetical patient with a "true" PHQ-9 of 12 (the midpoint of the moderately depressed category) the CV_T from the data in this study is 0.44. This suggests a probability of 0.57 that the measured result will be 3 points lower or 4 points higher, that is that the patient would be classified in a different category (mild or moderately severe).

It is therefore clear that the high level of between-day within-individual variation seen here would make measurements taken on a single day potentially unrepresentative of a patient's true status.

Clinicians should therefore be aware of the possibility for miscategorisation of depression based on the symptom scores derived from depression symptom measurement instruments, and the potential that this could cause inappropriate management of depression. Authors of clinical guidelines should take variation into account when making diagnosis, monitoring and treatment recommendations and researchers should also take this variation into account when assessing a significant change in score when evaluating a new treatment.

Note that some underlying conditions for which the tests considered in this thesis are performed have higher variability than others, that is the pathological component

of the total variation will comprise a greater proportion. For example, acute illnesses measured by CRP would likely have a higher pathological variation than depression.

Further research that addresses confounding by a change in frequency or timing of testing with increased disease severity or variation would be useful. Studies into how variation in this population affects clinical outcomes would also be important. Further investigations to disentangle the effect of variables in subgroups is needed using multivariable analysis. This may help answer questions arising from this study such as: are geographical differences in variation associated with differences in mean scores; is the lower variation associated with antidepressant treatment a physiological function, or related to lower mean scores; is the higher variation associated with South Asian and Black ethnicities associated with lower socioeconomic status, or could it be a function of the measurement tool in that it was validated for a different population? There is also a need to develop statistical methods and/or other scoring systems to evaluate the true variation of depression symptoms. In general there is a need to investigate the extent to which clinical decisions may be influenced by chance variation.

9.7 Conclusions

This study is by far the largest description of total measured within-individual variation of the depression symptom measurement instrument PHQ-9, and shows that there is considerable within-individual variation. However, in contrast to the other cohort studies in this thesis, the total measured variability was found to decrease as mean and median score increased for all three depression measurement

instruments. This is likely to be an intrinsic issue related to the design of the instruments.

In the final chapter of this thesis, I summarise the findings and discuss the implications of this research and make suggestions for further research.

10 Chapter ten: Discussion

10.1 Main findings: summary and implications for practice

10.1.1 Overall summary

The findings of this thesis show that:

1. the published data on within-individual variation for various clinical measurements is either poor-quality, for example small sample sizes or numbers of repeat measurements, poor in quantity i.e. few or no published studies, and/or poor in generalisability. Generalisability problems are related to a. differences in population – for example many of the studies only related to healthy populations; b. differences in ways of performing measurements, such as idealised research conditions with highly controlled pre-analytical conditions compared to real-world results; c. differences in time intervals, for example only short intervals between measurements.

Many of the studies in the systematic reviews were not designed specifically to assess within-individual variation. Those that were tended to be laboratory or research-based settings, leading to a higher accuracy but poorer generalisability than studies that reported variation from real-world settings.

2. the within-individual variation of the clinical measurements in this thesis are much higher than previously reported.

By contrast to the findings of the systematic reviews, the estimates of variation in this thesis from the IMRD and CPRD databases are much more clinically relevant than those in the systematic reviews, based as they are on very large datasets, large numbers of repeat measurements and real-world data. The real-world nature of the data is a particular strength of this study. It is vital when considering the effects of within-individual variation on clinical decision-making, that the estimates of variation are based on the test results on which the decisions are made. This is the case with CPRD and IMRD data, in contrast to the majority of the published data on variation to date.

The analyses of the data in the IMRD and CPRD databases are pragmatic estimates from routinely collected data rather than controlled trials. There is no information in the data as to the purpose of the tests, which makes it hard to account for factors such as confounding by indication (for example where more tests are done when a patient is more severely ill).

Nevertheless, these studies show that the within-individual variation in these large real-world datasets is much higher than previously reported. Even with the potential for bias and inaccuracy because of the real-world nature of the data, the size of the variation is so marked that I believe this has important clinical implications for clinicians and authors of clinical guidelines. Additionally, the within-individual variation for five of the eight measurands increases with increasing disease severity. The variation of the depression measurement instrument scores does not follow this pattern however, but this seems to be a function of the integer scale on which the results of these instruments are measured. A small change at the lower end of the scale will have a disproportionately higher effect on the CV than a similar change at

the higher end of the scale. This suggests there may be a problem with estimating variation in other PROMs.

Note that while there are marked differences in the extent of variation found in the cohort studies based on primary care EPRs compared to published data on variation in the systematic reviews, this is to be expected since the populations being investigated are different. Even with the higher quality studies (larger numbers of participants, larger numbers of repeat measurements), most of the previously published data is based on specific, often homogenous populations and usually with tightly controlled measurement conditions. The variation estimated in these studies will inevitably be lower than that calculated from real-world databases in which the population and the measurement conditions are not restricted and controlled. Nevertheless, since the real-world measurement results are the results on which clinical decisions are made, it is vital to know the extent of the variation in these real-world test results.

10.1.2 Systematic reviews

10.1.2.1 *Chapter two summary – HbA1c systematic review*

The systematic review of within-individual variation of HbA1c included 111 studies and a total of 564,143 participants (chapter two and [327]), which is much larger than the biggest previously published systematic review on the subject which included only 17 studies [132]. The median overall CV_T was 0.070 (IQR 0.034 to 0.090) which is much higher than the largest previously published review, which reported a CV_T of 0.013 (95%CI 0.012-0.025) [132]. However, this paper only reported results from studies in healthy subjects, and found a result similar to the

CV_T for the subgroup of healthy subjects in my systematic review. Twenty-five studies in my review described variation in healthy subjects, with a median CV_T of 0.017 (IQR 0.013 to 0.022); 19 studies described variation in subjects with type I diabetes, with a median CV_T of 0.084 (IQR 0.067 to 0.89); fifty-nine studies described variation in subjects with type II diabetes, with a median CV_T of 0.083 (IQR 0.060 to 0.100). It can be seen from this that participants with diabetes had much higher variation than participants without diabetes. This review also demonstrated a positive correlation between CV_T and mean HbA1c (Chapter two fig. 6) which had a R² value of 0.48. This finding has not previously been reported for HbA1c.

One limitation with the studies in this systematic review is that some of the studies did not record the method used to measure HbA1c (IFCC v NGSP). This can affect the calculation of coefficient of variation which is sensitive to transformations that involve adding or subtracting a constant [335].

10.1.2.2 Chapter three summary – CRP systematic review

The systematic review of within-individual variation of CRP is the first large scale systematic review on this subject (chapter 3), including 60 studies with a total of 75,293 participants. The only other compilation of CRP variation results (there is a meta-analysis but no systematic review) is on the EFLM database and includes at the time of writing five studies on CRP and two on hsCRP [77]. Thirty-four of the studies in the systematic review in this thesis reported results on CRP and twenty-six on hsCRP. Note that CRP and hsCRP both measure the same acute-phase protein, but the hsCRP is a high sensitivity test designed to detect small elevations, and is used primarily to help identify cardiovascular risk [159]. Thus the true within-

individual biological variation of CRP and hsCRP is identical, but the total within-individual variation will differ due to differences in the analytical component of the variation. The median CVs were indeed similar: for CRP 0.41 (range 0.11 to 0.89); for hsCRP 0.44 (range 0.27 to 0.76). This very high level of variation suggests that this test is highly inaccurate, although the usefulness of a test depends on how it is utilised.

It has been suggested that some studies on the biological variation of CRP are incorrectly high because of a failure to allow for the impact of transient illness on the results and/or because of non-normally distributed data [165]. These issues are discussed further in the section on the IMRD study of CRP variation (below).

Although this systematic review was much larger than anything comparable, the low numbers of studies on subjects with disease status other than healthy were too low to draw any conclusions about the effect of disease status on CRP variation.

The question of the extent to which CRP variation is a problem in clinical practice has not previously been answered, and it is of vital importance given the widespread use of this test for a variety of infectious and inflammatory diseases and as a biomarker for conditions such as cardiovascular disease and cancer [387].

10.1.2.3 Chapter four summary – Spirometry systematic review

The systematic review of within-individual variation of FEV1 and FVC included 104 studies with a total of 78,849 participants (chapter four). To the author's knowledge, this is the first systematic review on the subject of within-individual variation in spirometry. Forty-five (43.3%) studies were on healthy populations, 19 (18.3%) were on patients with asthma and 14 (13.5%) were on patients with COPD.

The overall median CV_T for FEV1 was 0.034 (IQR 0.025 to 0.067) and for FVC was 0.031 (0.02 to 0.048). In healthy populations, median CV_T of FEV1 was 0.028 (range 0.011 to 0.177) and of FVC was 0.029 (range 0.010 to 0.163). In COPD, the median CV_T of FEV1 was 0.068 (range 0.020 to 0.122) and the median CV_T of FVC was 0.068 (range 0.016 to 0.140). In asthma, the median CV_T of FEV1 was 0.030 (range 0.018 to 0.089) and of FVC was 0.024 (range 0.014 to 0.029). As with the other systematic reviews in this study, some studies that did not primarily aim to evaluate within-individual variation may have been missed. However, over a hundred studies on the subject were retrieved, showing there is a large body of data on the subject, albeit not necessarily of high quality or clinical relevance. Large cohort studies on spirometry that were not picked up by the search strategy were also considered, but despite having a large amount of longitudinal data, they did not report within-individual variation of spirometry measures [388, 389]. It is possible that a future analysis of the individual patient data of these studies would provide more data on this subject.

Some studies evaluated the within-day variation and some evaluated the between-day variation. Patient effort and technique is important in gaining accurate results, and so repeated within-session tests are routinely performed. The within-day variation can be viewed as the sum of the pre-analytical, analytical and within-individual same day biological variation. The between-day variation of spirometry measures in this review was slightly higher but similar to the within-day variation measures, suggesting that the within-day variation is a major source of the overall variation of spirometry measures. However, what constituted within-day variation differed by study protocol. For example, Abbas [207] performed two baseline tests of FEV1 fifteen minutes apart, while Alamer [271] and Baptist [272] performed three

tests on the same day but did not report the time interval between the tests and Connolly [211] performed two tests of FEV1 and FVC at least one hour apart.

Long-term assessment of variation of spirometry measures is complicated because of the fact that lung function declines over time. This is true of healthy individuals but is worse in patients with COPD, who may show a mean decline in FEV1 of 33-69mL/year [357]. However, most of the studies in this paper looked at a study period of less than a year, minimising this problem.

There is even less high-quality data on the within-individual between-day variation of spirometry measurements than for HbA1c and CRP, and it seems to be generally considered that because within-day variation is compensated for by repeated measurements within a session, within-individual variation is therefore not a problem. However, there is currently insufficient evidence to draw this conclusion.

10.1.2.4 Chapter five summary – Depression symptom scores systematic review

The systematic review of within-individual variation of depression symptom measurement instruments is to the author's knowledge the only systematic review on this subject (chapter 5 and [383]), including 41 studies and a total of 4,984 participants. Twenty-five studies were on PHQ-9, fourteen on HADS or HADS-depression subscale and seven on BDI. Only one study reported a CV_T , which was 0.42 for the PHQ-9 in patients with subclinical hypothyroidism [306]. The other studies reported the ICC (that estimates the ratio of variation) which was therefore the variation measure used in the systematic reviews. The pooled estimates for ICC

(with 95% CI) were for PHQ-9 0.84 (0.81-0.88), for HADS 0.89 (0.85-0.93) and BDI 0.89 (0.84-0.93).

However, the data on within-individual variation of these depression symptom measurement instruments was poor. Twelve studies were on healthy populations and only three on patients with mental health conditions, none of which was depression. This is important since it would be very useful to know how much variation in depression scores there is in depressed individuals in order to make recommendations for changes in treatment as their condition improves or worsens, e.g. the addition or withdrawal of anti-depressant drugs. Most studies involved two measurements and a time interval of two weeks or less, and most were performed in order to assess test-retest reliability rather than within-individual variation. The low numbers of repeat measurements limits the accuracy of the variation estimates, and the short time-interval between tests means there is a lack of data on the long-term variation of these tests. Sample sizes were fairly small, with a range of 15 to 281 participants (median 70). It is currently hard to draw any conclusions about the within-individual variation of depression measurement symptoms given the almost complete lack of useful published data on this subject.

10.1.2.5 Overall summary of systematic reviews

Systematic reviews were performed to assess the published data on within-individual variation in seven measurements that are performed commonly in general practice: HbA1c; CRP; the spirometry measurements FEV1 and FVC; depression symptom measurement instruments PHQ-9, BDI and HADS. In total, 8,724 abstracts were screened, 873 full texts were examined and 316 studies were included in these

four systematic reviews. Each of the reviews concluded that many of the included studies had low numbers of participants and/or low numbers of repeated measurements, reducing the accuracy of the variation estimates (see strengths and weaknesses of systematic reviews below). It was also noted for HbA1c, CRP and spirometry that participants with the disease for which the measurement was being performed had higher within-individual variation than healthy participants. This data was not available for the depression symptom measurement scores. The CV_{TS} from the systematic reviews are given in table 10.1, together with details of the number of studies, participants and measurements.

For the HbA1c review, it was seen that higher HbA1c levels and diabetic status were associated with higher within-individual variation, but it was hard to draw similar conclusions from the other systematic reviews due to limited data on patients with the disease and/or low numbers of studies and patients.

Table 10.1 Median CVTs of systematic reviews.

* ICC with 95%CI.

Measurand	Number of studies	Range of number of participants	Total number of participants	Range of number of measurements	Median overall CV _T (IQR)
HbA1c	111	4 to 91,866	564,143	2 to 37	0.070 (0.034 to 0.090)
CRP/hsCRP	60	4 to 56,218	75,239	2 to 18	
• CRP	34	-	-	-	0.41 (0.11 to 0.89)
• hsCRP	26	-	-	-	0.44 (0.27 to 0.76)
Spirometry	104	4 to 26,368	78,849	2 to 837	-
• FEV1	97	-	-	-	0.034 (0.025 to 0.067)
• FVC	59	-	-	-	0.031 (0.020 to 0.048)
Depression scores	41	15 to 458	4,984	2 to 13	-
• PHQ-9	25	-	-	-	0.84 (0.81–0.88)*
• BDI	7	-	-	-	0.89 (0.84–0.93)*
• HADS	14	-	-	-	0.89 (0.85–0.93)*

10.1.2.6 *Strengths and limitations of Systematic reviews*

The strengths of the systematic reviews are that they all represent the latest and largest systematic reviews on the subjects of within-individual variation on the measurands they investigate. In fact, for spirometry and depression symptom measurement instruments, these are the only systematic reviews on the subject.

There are some weaknesses common to all the systematic reviews, which can be broadly categorised as weaknesses with the review process and weaknesses with the published data.

For the review process, no suitable risk of bias tool was available. I made the decision to use a validated risk of bias tool (the COSMIN tool) rather than the bespoke tool I had designed and included in the protocol. The COSMIN tool has some limitations, particularly when it comes to the questions regarding blinding. As discussed in Chapter 2.6, generally in studies into biological variation the investigators are not blinded to previous measurements. These questions were therefore omitted when a study's risk of bias score was assigned. This is unlikely to have affected findings for laboratory tests since they are measured objectively. For spirometry, knowledge of previous scores and assignment of scores may have some influence on the decision of the technician performing the test to repeat the test or be content that the blows are acceptable – a result unexpectedly different from a previous result might prompt the technician to ask the patient to perform further blows. For the depression symptom measurement instruments, the patients themselves may recall their previous answers, especially when the tests are performed at short intervals. In all the tests considered in this thesis, knowledge of

previous test scores might influence the decision to perform the test in the first place, which could lead to confounding by indication. Another possible limitation is the use of a scoring system, which restricts the risk of bias assessment to limited questions and results, rather than a narrative risk of bias assessment which could give a more detailed appraisal of the risk of bias of each study.

A third risk of bias tool. The BIVAC instrument was considered [154]. This has been designed to assess risk of bias in studies of biological variation, but only considers the CV as the metric of variation, which would exclude other measures of variation. Further, it is only suitable for laboratory tests and so I chose to utilise a risk of bias tool that was relevant to all four systematic reviews.

The search strategies were designed to be as broad as possible, but since measuring variation was often not the main purpose of the studies, it is possible that data on within-individual variation within the main text of certain studies was not picked up. Despite this, these four systematic reviews are the biggest or only reviews on their subjects. No validated method for performing a meta-analysis of CVs has yet been described, so reports of the CVs had to rely only on descriptive statistics. The EFLM database uses a weighted median method for global CV estimates, but the method does not appear to be well-described or validated.

For the published data itself, there were some common weaknesses. Many studies had missing data such as ethnicity. Many of the published studies included small number of participants and/or small number of repeated measurements, often with small time intervals between measurements. Note that although some of the studies had large numbers of participants or large numbers of measurements, the studies that had large numbers of measurements tended to have low number of participants and vice versa. Many of the studies were performed under idealised conditions that are poorly generalisable to clinical practice, with few utilising the real-world data that clinicians use to make clinical decisions.

Another common problem was that it was also not entirely clear whether biological variation or total variation (within-individual biological plus analytical variation) was being described in many cases. For the main analysis all variation measures were included, but for the HbA1c review there were sufficient studies to perform a subgroup analysis with results grouped by whether total or biological variation was described or whether this was uncertain. This showed a larger variation in studies describing total variation than biological, which is expected given total variation is biological plus analytical variation, and the studies describing total variation are often real-world studies with less tightly controlled test conditions than those describing within-individual biological variation.

10.1.3 IMRD and CPRD database studies

10.1.3.1 Chapter six summary – IMRD HbA1c study

The within-individual variation of HbA1c was estimated from the IMRD database. 587,023 participants that had at least four measures of HbA1c recorded as mmol/mol with values between 20.1 (4%) and 195.1 (20%) mmol/mol were included in the study, more than all the participants in the studies in the recent systematic review on this subject put together. The overall total within-individual CV_T was 0.20 (95%CI 0.20 to 0.20). This is around three times higher than reported in the systematic review, which had an overall CV_T of 0.07. The CV_T for non-diabetic patients was approximately five times higher than in the systematic review, and for diabetic patients approximately twice as high.

CV_T increased with median HbA1c up to values which would be considered indicative of poor control of diabetes. As with other measurements in this study (apart from depression symptom measurement instruments), a result indicating a more severe disease status was associated with a higher variation. Diabetic status and the use of anti-hyperglycaemic drugs were also associated with higher variation. This is a scientifically plausible finding, since patients with higher HbA1c are likely to have poorer intrinsic control of their glucose levels due to their low insulin levels, and have more variation introduced into their glucose levels by their medication regimes. However, it is an important finding for clinicians to be aware of, since the measurements on which treatment decisions of the most severely ill patients are taken are the least likely to reflect the true mean status of the patient.

Possible weaknesses associated with this study such as errors in data entry were minimised by removing implausible values, and in any case these errors were likely to be low since the values are automatically entered into the system. Bias was reduced by initially including all patients in the database, with the main restriction

being those that had less than four measurements. On sensitivity analysis, lower number of measurements was associated with a lower variation, but this is likely due to more poorly controlled patients having a larger number of measurements performed.

10.1.3.2 Chapter seven summary – IMRD CRP study

The within-individual variation of CRP was estimated from the IMRD database. 472,593 participants that had at least four measures of CRP recorded as mg/L between 0 and 1,000 were included in the study, which is five times the number of the sum of the participants in all the studies identified in the systematic review of CRP variation (See Chapter three). The overall CV_T was 1.148 (95% CI 1.147 to 1.149), which is much higher than the median reported for previous studies of 0.41 (range 0.11 to 0.89). When looking at subgroup analyses, median CRP level had the largest effect on CV_T , with a marked increase above a median CRP of 10-20mg/L, approximately the level where CRP elevations are considered clinically important. As was the case for HbA1c, participants with a result indicating a more severe status had a higher within-individual variation.

The results of within-individual variation are more challenging to interpret than the HbA1c results. Firstly, CRP has a short half-life, and so transient illness can affect the apparent variation. This can be seen in the sensitivity analysis in which much higher variation is seen in participants with a very short median interval between measurements (< 14 days). However, these participants only accounted for around 15% of the total participants in the study, and the CV for those with longer median intervals between measurements was still always above 1.00.

Another difficulty with CRP is the highly-skewed nature of the data, since normal ranges are in single figures, moderate disease severity is associated with ranges in double digits and severe disease is associated with ranges in triple figures. This invalidates the statistical methods used for the other measurands in this thesis which had normally distributed data. To overcome this, log-transformed data was used. Log-transforming data does not guarantee normality [165]. Statistical tests such as Shapiro-Wilk's W test for normality are not valid at such large values of N, but a visual examination of the log-transformed data showed it was approximately normally distributed, making the use of the mixed effects linear regression method for calculating the CV_T valid.

Finally, the IMRD database only lists CRP as a test and does not distinguish CRP from hsCRP. There appears to be little data on how widespread the use of hsCRP is in primary practice, but since there seems to be a reluctance to utilise this test [390] and it is not currently recommended by NICE as a screening test for cardiovascular disease, it is likely that the vast majority of the tests recorded in the database are CRP and not hsCRP. Even if this was not the case, the systematic review shows that the within-individual variabilities for CRP and hsCRP are similar and so mixing recordings of hsCRP and CRP in the database should have little impact on the results and conclusions of this study.

It should be noted that the ranges for diagnostic treatment decisions in CRP are reasonably wide, but despite this the within-individual variation of CRP is so high that single measurements of CRP will frequently miscategorise patients, for example when deciding whether to treat respiratory infections with antibiotics. Clinicians and guideline authors need to take the high within-individual variation of CRP into account when making clinical decisions or writing guidelines.

10.1.3.3 Chapter eight summary – IMRD spirometry study

The within-individual variation of FEV1 and FVC was estimated from the IMRD database. 4,412 and 3,567 participants with at least four measures within a six-month period and plausible values were included for FEV1 and FVC respectively. The overall coefficient of variation for FEV1 was 0.224 (95% CI 0.221 to 0.228) and for FVC was 0.152 (0.150 to 0.155). This is much higher than seen in the systematic review of spirometry variation (Chapter four) which reported a median CV for FEV1 of 0.034 for FVC of 0.031. This may be because of the fact that nearly half of the studies were in healthy populations which had lower variation than those with respiratory disease, and because nearly three-quarters of the studies were conducted in hospitals, laboratories and other secondary or tertiary care settings which are likely to have more tightly controlled pre-analytical and analytical methods than in primary care.

BMI, age, smoking status and respiratory diagnosis had an effect on variation in univariable analysis. CV_T decreased with increasing values of FEV1 and FVC, that is lower values were associated with higher within-individual variation. This is an interesting finding that helps validate the findings in the HbA1c and CRP studies that variation increases with disease severity. In this case, unlike with HbA1c and CRP, more severe disease is associated with lower values of FEV1 and FVC, and this shows that the association between variation and disease status is a real finding and not simply a mathematical function of CV being for example proportional to the mean.

There are some difficulties with interpreting variation of spirometry measures. The most important is distinguishing true variation from decline in lung function, which happens in everyone as they age and is worse in COPD. The CV does not distinguish between long-term trends and chance variation. For example, if successive hypothetical FEV1 values showed a fluctuation around a mean of 1, 5, 4, 2 and 3, the CV would be 0.11, which is the same as if the FEV1 values showed a systematic decline of 5, 4, 3, 2, 1. Spirometry is the only one of the four primary care database studies that is affected by this issue, since HbA1c, CRP and depression symptom measures would not be expected to worsen (i.e. indicate a decline in health) over long time periods, since diabetes, inflammatory diseases and depression are not generally progressive diseases if they are controlled adequately. To compensate for this issue, the variation calculations were performed only on measurements performed within six months of the first measurement, a time period which should be short enough that decline in function does not have a significant effect on the results. This limits the amount of subgroup analysis that can be performed due to the lower numbers in each subgroup, although it can be seen that in sensitivity analyses by mean FEV1 and FVC results and by diagnosis in which all measurements were included irrespective of time, there was little difference compared to the short-term analysis.

Nevertheless, there were sufficient participants to show that respiratory diagnosis had an important effect on spirometry variation, with a particularly large increase in FEV1 variation in participants with COPD. Clinicians are aware of the within-day variation of spirometry measures and make efforts to minimise this by for example repeating measurements in a session. However, literature searches and my personal communication with respiratory clinicians suggest that the extent and effect

of between-day variation of spirometry measures is not fully appreciated. This could lead to errors in diagnosis and management of respiratory diseases such as COPD and asthma.

10.1.3.4 Chapter nine summary – CPRD depression symptom score study

The within-individual variation of depression measurement symptom scores was estimated from the CPRD Aurum database, since these scores are not recorded in the IMRD database. Both databases record similar patient data, though the CPRD database is larger than the IMRD database [391] (the CPRD database used for this thesis included 24,176,224 participants registered in 1491 practices compared to 11,737,653 participants registered in 832 practices for IMRD). 195,171 participants had at least four measures of PHQ-9 after impossible values were excluded. For HADS, this figure was 8,231 participants and for BDI 215. Most of the subgroup analyses were therefore only performed on the participants with PHQ-9 scores recorded. The overall CV_T (95% confidence interval) was 0.412 (0.411 to 0.413) for PHQ-9, 0.370 (0.365 to 0.374) for HADS and 0.763 (0.701 to 0.826) for BDI. Only 5,116 participants were included in all the studies in the only systematic review on this subject (Chapter five and [383]), and in any case only one of those studies (Karmisholt [306]) reported a CV of a depression measurement instrument (for HADS with a CV of 0.416 in fifteen patients with subclinical hypothyroidism). The study in this thesis is the only major study of within-individual variation of depression symptom scores and demonstrates a large within-individual variation for these scores. It is more common for laboratory tests to have their within-individual variation measured, often expressed as a coefficient of variation since this value is used to set analytical precision standards. Symptom scales do not usually have their variation assessed in this way.

However, unlike for HbA1c, CRP, FEV1 and FVC, variation decreased with results representing increasing disease severity for PHQ-9, HADS and BDI. This is a surprising result, since it is contrary to the results of the other studies, and seems counter-intuitive, since it would be expected that patients with more severe depression would have a larger fluctuation in disease severity than patients without depression, whereas the results suggest that patients with severe depression are consistently depressed while those with mild or no depression vary much more. The explanation likely lies with the integer scale used in these and other PROMs. Since the scale is not continuous, small variations at the lower end of the scale will have a larger effect than variations at the higher end of the scale. On a continuous scale, small relative changes in a measurement are possible at all values. A result of 2 may increase or decrease by for example 0.2, a 10% increase. However, on a scale in which only integers are possible, for a result of 2, the minimum changes possible are + or -1, a change of + or -50%. For a result of 20, a change of + or -1 would be only a 5% change. This effect is much larger than any other effects of variables such as diagnosis of depression or drug use, meaning that patients with a diagnosis of depression or being treated with anti-depressants had a lower variation than those without a depression diagnosis or anti-depressant treatment, despite the fact that their subjectively experienced symptoms may in fact be more variable than the depression measurement scores indicate.

This finding has an important implication for clinicians and clinical guideline authors. Results at the lower end of a depression measurement symptom score scale are less likely to represent the true status of the patient, which might mean that estimates of depression severity in the mild to moderate range are inaccurate.

10.1.3.5 Overall summary of IMRD and CPRD database studies

Within-individual variation was estimated from the IMRD and CPRD databases for the seven measurands studied in this thesis using a mixed-effects linear regression model (see Table 10.2). The within-individual variation in these studies is the total within-individual variation, including the pre-analytic, analytic and biological variation. This is the most useful measurement of variation for clinicians, since it reflects the data they use for making clinical decisions.

Table 10.2 Summary of IMRD/CPRD studies of variation.

Measurand	N	Median number of measurements per individual (IQR)	Overall CV _T (95% CI)
HbA1c	587,023	11 (7-18)	0.200 (0.200 to 0.200)
CRP	472,593	10 (6-32)	1.148 (1.147 to 1.149)*
FEV1	129,161 (4412)	8 (5-11)	0.224 (0.221 to 0.228)
FVC	75,956 (3567)	7 (5-10)	0.152 (0.15 to 0.155)
PHQ-9	195,171	5 (4-7)	0.412 (0.411 to 0.413)
HADS	8,231	6 (5-8)	0.370 (0.365 to 0.374)
BDI	215	5 (4-7)	0.763 (0.701 to 0.826)

*N is number of participants used for data analysis i.e. after removal of implausible values, with at least four measurements. For spirometry data the number in brackets is number of participants used for analysis where measurements were within six months of the initial measurement. *CRP results calculated from log-adjusted data.*

The CV_Ts were invariably substantially higher than those reported in the literature. For measurands on a continuous scale (HbA1c, CRP, FEV1, FVC), there was a consistent association between more severe disease status and higher within-individual variation. This held, regardless of the direction in which the result indicated disease severity. That is, for HbA1c and CRP, higher values indicate more severe disease, whereas for FEV1 and FVC, lower values indicate more severe disease. The finding of a higher variation with a more severe disease is important, because while it is usually assumed that the coefficient of variation is independent of the mean, this isn't always the case [392]. The fact that this relationship holds regardless

of whether severity is indicated by an ascending or descending scale shows that the finding that within-individual variation worsens with disease severity is a real finding and not a mathematical quirk.

This finding is in fact in my opinion consistent with clinical experience and basic pathophysiology. Through a variety of mechanisms, mostly involving negative feedback loops, the body tries to keep its internal environment in a constant state, the concept of homeostasis. Disease tends to alter the body's homeostatic mechanisms, either deliberately (e.g. increased acute phase proteins in response to infection), or through the disease itself (e.g. increased glucose due to relative or absolute insulin deficiency), which will reduce the body's ability to maintain that value within as narrow a range as when in a healthy state, i.e. variation will increase. Fluctuation in the disease state itself (e.g. asthma severity varying with the level of air pollution) and in individuals' response to the disease state (e.g. altering nutritional intake and activity levels, taking medication) will also contribute to increased variation in disease compared to a healthy status.

However, for all three depression symptom measurement instruments there was a consistent marked negative association between average score and variation, with the effect particularly pronounced at the lowest end of the scale, that is where the average values were nearest to the healthy state. This is likely due to the mathematical properties of the scales rather than a true association between disease status and variation (see discussion on depression symptom measurement instruments above). This effect was much larger than the effects of age, medication usage or depression diagnosis on variation.

This has important implications when considering the reliability of depression symptom measurement instruments, and by extension all PROMs that utilise integer scales, with the most important conclusion being that lower values may be highly unreliable due to the extent of the variation intrinsic to the scale. This may have little importance in healthy individuals where a treatment decision is not necessary, but may affect decision-making in the mildly to moderately depressed. Unfortunately, there has been little work published on methods of evaluating the variation of categorical and integer scales [393].

10.1.3.6 Strengths and limitation of IMRD/CPRD studies

Certain strengths and weaknesses are common to all the primary care database studies of within-individual variation in this study.

The two major strengths are the large number of participants and the real-world nature of the data used. All the studies in this thesis had many more participants than any previous studies, and in fact each study in this thesis had more participants than all the participants in the systematic reviews put together. This allowed some subgroup analysis to be carried out, to assess the effects of demographics, diagnoses and treatments on within-individual variation.

Although the real-world data could be viewed as a weakness of this study, in that the measurements are not performed under strictly controlled conditions to minimise pre-analytical variation, this is in fact a strength when it comes to utilising the results of these studies for the purpose of improving clinical decision-making and clinical guidelines (as opposed to setting laboratory analytical standards). The data in the IMRD and CPRD databases are the results used in clinical practice, and

therefore it is important to evaluate the variation of the data that clinicians use to make clinical decisions. Real-world data is in fact being increasingly used to make clinical guidelines and to augment data from randomised controlled trials [394].

There are some weaknesses with the use of big data from Electronic Patient Records (EPRs). It can be considered “dirty” in that it relies on accurate input from busy medical practitioners, and there is a lack of standardisation of methods. Observational studies using EPRs might be prone to bias that is reduced in randomised clinical trials. For example, the date of recorded diagnosis is not necessarily the date when the disease began. The clinician’s decision to order a test introduces selection bias (see below). The timing of data entry can be affected by weekends, holidays and efficiency of administrative processes [338]. Data may also be missing, and this data may not be missing at random (for example, ethnicity may be less likely to be recorded in predominantly white populations, and it is of note that in the studies in this thesis, “missing” ethnicity often has a within-individual variation closest to “white” ethnicity).

Some factors minimise these weaknesses in this thesis. For HbA1c and CRP, the results are transferred automatically by the laboratory, removing the possibility of input error by clinicians. However, it is clear that errors are frequently made in the THIN and CPRD data, for example selecting the wrong units for results. Data cleaning (for example removing impossible or implausible values) has helped to remove many of these errors. The initial inclusion of the entire database reduces the risk of bias, but excluding participants with low numbers of repeat measurements may introduce some confounding by indication, with more severely affected patients having larger numbers of repeat measurements.

Possible problems with transient illness (CRP), non-normal data (CRP) and long-term decline (spirometry) were discussed in the individual studies.

There are a large number of statistical methods of calculating variation, and several were considered. Standard deviation was not used because of its dependence on the mean. Various statistical methods that produce a variation independent of the mean (VIM) have been described. However, VIM is a statistical tool, not a clinical measure and does not generalise to other datasets [395]. CV is a traditional and well-recognised measure of variation and was therefore selected as the main variation measure for this thesis.

Distinguishing between a change in results due to a material change in the underlying disease as opposed to baseline biological variability or increase variability of measured results due to the disease process (a wider variation around the homeostatic set point caused by pathology) is difficult in electronic patient records without clinical notes to summarise the disease status of the individual taking into account all clinical findings.

There is a risk of bias in these studies due to the fact that patients undergoing tests are considered at risk for the disease in question, and this is accentuated by the requirement in these studies for patients to have at least four repeat measures, suggesting that even those not diagnosed with a disease are considered at high risk for the disease. It has been shown that the presence of a laboratory test order in the clinical records, regardless of any other information about the test results, has a significant association with survival in 86% of tests, an example of selection bias which could have implications for the results of this study [338]. Since even those patients classified as healthy in this study are likely to be at higher risk of the target

disease for which they are being tested, this might partly explain the higher level of variation in the “healthy” populations in this study compared to that seen in controlled prospective biological variation studies on healthy subjects. The results of this study should therefore be interpreted somewhat cautiously in the light of the possible overestimation of variation caused by selection bias in the EPRs.

There is a need for multivariable analysis to clarify the effects of long- vs short-term variation and interaction between other influences on variation such as socio-demographic factors and disease status.

Overall, the strengths of the studies in this thesis, namely the very large sample sizes and the real-world nature of the data, in my opinion greatly outweigh any weaknesses in the data.

10.2 Comparison to other research

10.2.1 Systematic reviews in this thesis

As described above and in chapters two to five, some work has been previously done on the within-individual variation of the measurands investigated in this thesis, although there are limitations with the published data which is discussed above. However, there is some research into the individual variation of other clinical measurands, with the variation of cholesterol and blood pressure being particularly well-described.

10.2.2 Analogous studies

Various estimates of within-individual variation of blood pressure have been reported and it is well-recognised that blood pressure variation has prognostic implications, leading to several systematic reviews of blood pressure variation and its

effect on outcome in a wide range of conditions such as cardiovascular disease, dementia and stroke [76, 396, 397]. Some research has also been done into factors that affect blood pressure variation, such as thiazide diuretics [398]. Research into blood pressure variation has some similarities with the work in this thesis, since it usually concerns itself with the total measurement variation of the results clinicians use to make decision, and not with components of variation (analytical versus biological). As with some of the examples discussed in this thesis, it has been shown that within-individual variation of blood pressure measurements can lead to misdiagnosis [43], and even when repeated measurements are performed at home, nearly three-quarters of patients with a mean blood pressure of 120-130mmHg might be incorrectly diagnosed with hypertension [399].

There have been a large number of studies on the within-individual variation of a large number of laboratory tests, mainly with the purpose of ensuring optimal levels of analytical precision, with the EFLM database maintaining data on the variation of a large number of biochemical and haematological tests (see current initiatives below). The variation of some laboratory tests have also been analysed further in regard to the clinical implications of within-individual variation. The within-individual variation of cholesterol is well-described, and has been shown to be associated with outcomes such as cardiovascular mortality [400] and renal function decline [401].

10.3 Applications of Research

10.3.1 Current initiatives

The concept of biological variation was particularly brought to the attention of the scientific medical community by Fraser's textbook and other work on the subject [15], although published literature on the problem goes back as far as 1934 with Snedocor's discussion of the issue using examples from agricultural botany [402]. Previous databases and diverse sources of data on biological variation were consolidated after the 1st European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Strategic Conference into the EFLM biological variation database [77, 403]. At the time of writing, the EFLM biological variation database has 187 meta-analysis derived estimates of biological variation and over 3,000 individual biological variation estimates. To be included in the database, studies are assessed with the Biological Variation Data Critical Appraisal Checklist (BIVAC) [154]. This rates studies on fourteen questions scored from A to D on factors such as study design, measurement procedure and statistical methods. The overall score is the lowest score awarded for these questions, and those studies with a score of D are not considered suitable for inclusion in the database.

Note that the cohort studies on variation included in this thesis would likely be graded C by the BIVAC checklist due to issues such as lack of information on pre-analytical methods to minimise variation and on the analytical methods used. However, the studies in this thesis are based on the real-world data that clinicians use to make clinical decisions (see Strengths and Limitations).

The EFLM is a vital resource for clinical laboratories, since a major use of biological variation is for setting analytical performance standards with respect to imprecision, bias, total error and measurement uncertainty [403]. It is generally held

that for a test to be useful, its maximum imprecision should be less than half the CV_i [32]. However, one problem with its generalisability is that because it is skewed towards publications in strictly controlled and well-described circumstances (e.g. pre-analytical procedures, analytical methods, steady state nature of population) it is biased away from estimates of variation based on real-world datasets. A second problem is that it only holds data on clinico-pathological tests i.e. haematology and biochemistry. Tests based on physical/physiological measurements (e.g. weight, blood pressure, spirometry) or on PROMs and questionnaires (e.g. pain scores, dementia and depression symptom measurement instruments) are not included.

It should also be reiterated that biological variation is only a part of the total measurement variation seen by clinicians, which also includes pre-analytical and analytical variation. As such, splitting total variation into biological and analytical variation is mainly of interest to laboratory scientists. However, distinguishing between biological variation and pathological variation is important in order to filter the "signal" from the "noise" of measurements. Identifying how much variation is due to random variation biological variation, how much is due to increased variability in disease, and how much apparent variation is due to material changes in the underlying disease would help clinicians make more informed diagnostic and treatment decisions, by being able to better respond to real changes in the underlying disease process, and to better measure the response of interventions.

There are no current initiatives for a database of the variation of non-laboratory clinical tests.

10.4 Contribution of the research findings

The impact of within-individual variation on clinical practice is related to the purposes for which clinical measurements are performed. Broadly, measurements are performed in clinical settings for one of three reasons:

1. To aid in making a diagnosis
2. To provide information on prognosis
3. To monitor disease and make recommendations for starting or adjusting treatment

10.4.1.1 *Diagnosis and prognosis*

The effect of within-individual variation on diagnosis and prognosis can be considered together since similar considerations apply. The most important problem to which within-individual variation contributes is making a clinical error. Treatment recommendations and prognostic predictions are based upon the results of large-scale studies in which treatment effects and data for aiding prognosis are averaged across large numbers of people. However, applying these recommendations or predictions to an individual can be erroneous if a measurement performed in the individual does not represent that individual's true mean state. How far a single measurement is likely to be from the true mean is estimated by measures of within-individual variation such as the CV.

The likelihood of an error in diagnosis because of within-individual variation is in part related to the extent of that variation. However, it also relates to the threshold for diagnoses and the normal range for a measurement for healthy individuals and individuals with disease. A measurand which has a high variation but a large

disparity in ranges between diseased and healthy individuals might be less prone to error in diagnosis. For example, healthy individuals tend to have CRP levels less than 3mg/L. NICE guidelines recommend giving immediate antibiotics to patients presenting with acute respiratory symptoms that have a CRP >100mg/L, on the basis that these are most likely to have a bacterial pneumonia [404]. Even the large within-individual variation of CRP is unlikely to lead to incorrectly categorising a healthy person as having bacterial pneumonia. However, problems in diagnosis are more likely to occur near the boundaries between diagnostic or treatment categories. For example, someone with a true mean CRP of 110mg/L who should be given immediate antibiotics would commonly have a measured CRP of <100mg/L and so would be given only backup antibiotics.

Note that reference ranges and decision-making thresholds are similar but not identical concepts as discussed in chapter 1.7.

Repeated testing may reduce the likelihood of incorrectly diagnosing diseases, although it cannot completely eliminate the problem. However, repeat testing can lead to other errors if a result is not considered in the context of previous measurements. Repeated screening tests will inevitably lead to an erroneous diagnosis if performed enough times, if only the results of the single test are taken into account. For example, in a simulation of repeat screening of HbA1c in 100,000 people according to guidelines, between 1,602 and 2,233 people with true values in the normal range would be incorrectly diagnosed with pre-diabetes after three screens and a further 627 to 1,672 people with true values in the prediabetes range would be incorrectly diagnosed with diabetes after five screening tests [405]. The true situation may be much worse than this, since that study used a total CV of 0.036, whereas the CV for HbA1c from the systematic review of HbA1c variation in

this thesis was 0.070 (Chapter 2 and [327]) and the CV estimated from the IMRD database was 0.20 (Chapter six). The authors of that study recommended that equivocal results should undergo repeat testing, and that decisions are made on the average of results rather than single measurements.

This is particularly important for those diseases which have a unidirectional diagnostic pathway, i.e. a diagnosis is permanent once it has been made. A patient whose CRP returns to normal after acquiring bacterial pneumonia may be considered cured. But patients that are diagnosed with diabetes or depression who have an improvement in symptoms and measurements are not considered to be healthy, but are categorised as patients in remission from diabetes or with a history of depression, which will alter the way they are monitored and treated in the future. A medical history of certain conditions such as hypertension or mental illness can also affect an individual's life in a number of non-medical matters, from the cost of life insurance to how one is viewed in the criminal system. For example prisoners with a mental illness are likely to serve longer sentences than those without [406].

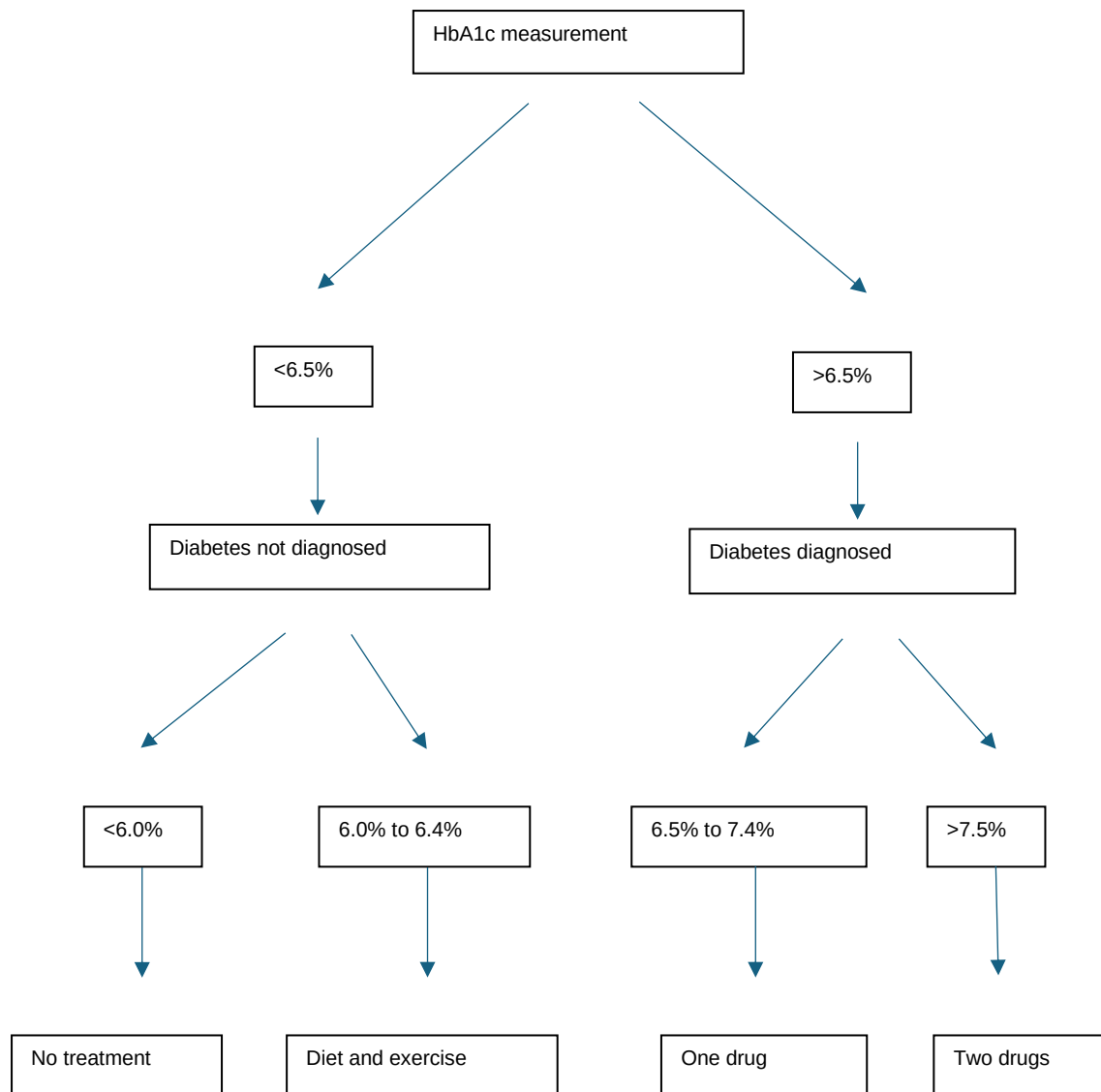


Figure 10.1 Simplified decision tree for diagnosis and management of diabetes based on NICE guidelines[339].

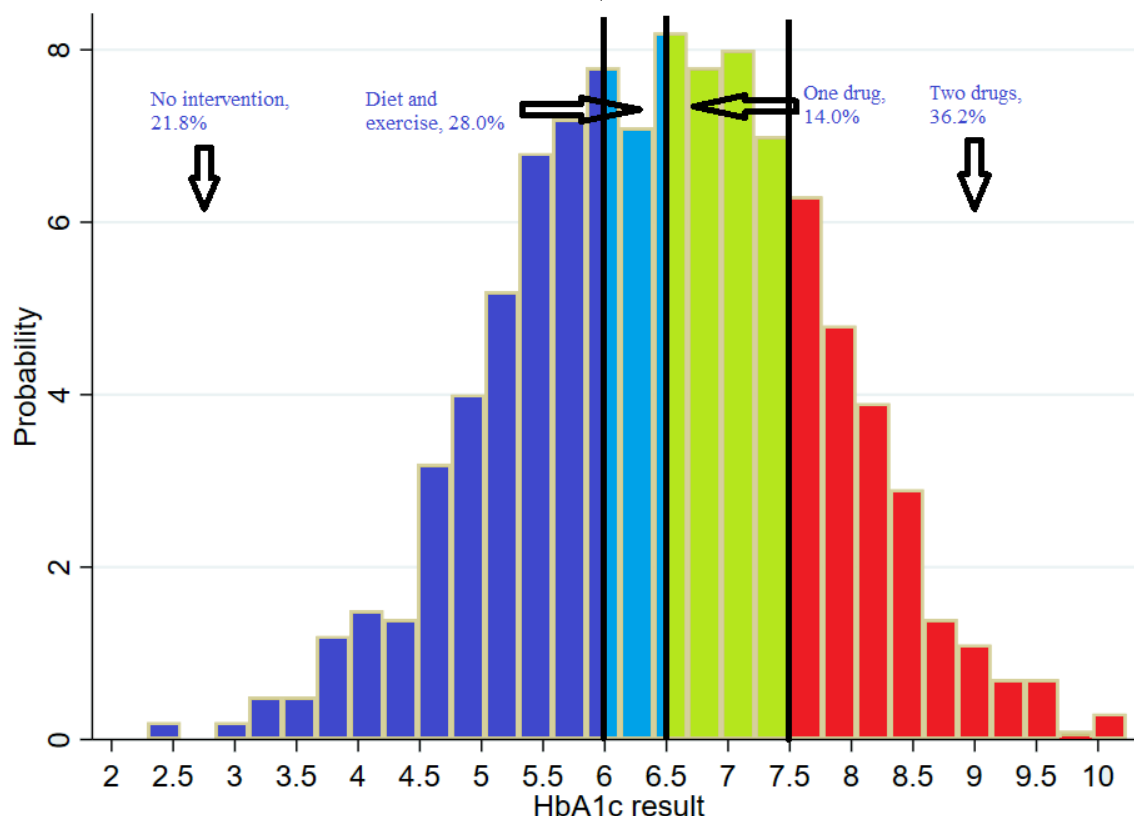


Figure 10.2 Simulation of distribution of 1000 patients with a true mean HbA1c of 6.5% (48mmol/mol) and a CV of 0.20.

HbA1c result on the x axis is in % and probability of this result being the measured finding is on the y axis in %.

Fig 10.1 shows a simplified decision tree for the management of diabetes mellitus based on current NICE guidelines. NICE guidelines recommend a single HbA1c measurement in patients with appropriate symptoms, although they state that a repeat test is “sensible.” [328]. In asymptomatic patients they recommend a second HbA1c measurement. The intensity of recommended treatment is based on HbA1c levels [339]. Fig 10.2 shows simulated data for the measured values of HbA1c in 1000 patients with a true mean HbA1c of 6.5% (48mmol/mol) and a CVi of 0.20 (the overall figure for HbA1c variability estimated from IMRD data in chapter six). It can be seen from this that based on a single measurement, the same patient

would have a 21.8% chance of being diagnosed with no diabetes and recommended no treatment, a 28.0% chances of being diagnosed with prediabetes and recommended diet and exercise, a 14.0% chance of being diagnosed with diabetes and recommended a single drug and a 36.2% chance of being diagnosed with diabetes and recommended two drugs (although in practice it is likely that a response to a single drug would be assessed before adding a second drug).

10.4.1.2 Monitoring

Clinical measurements are commonly used for monitoring chronic conditions such as COPD and diabetes. Clinicians may use repeated measurements within an individual to make inferences about within-individual effects in order to personalise their medical treatment. However, the within-individual variation of the measurements may make the recommendations to achieve the desired within-individual effects unreliable [407]. For example, if we take a hypothetical patient with a true mean HbA1c of 6.5% who is undergoing an HbA1c test for the purposes of monitoring rather than initial diagnosis, we can see from fig 10.2 that the probability of an HbA1c measured value leading to the most intense treatment recommendation of two drugs is about a third, and a recommendation of no intervention at all is about a fifth. These findings about the potential for misclassification of diabetes in both diagnosis and monitoring are consistent with a recently described model of misclassification of diabetes using HbA1c [408].

In the past, conclusions as to whether a treatment is effective have revolved around whether there is a statistical difference in effect between two treatments, or treatment and placebo. In 1989, the concept of minimal clinically important difference

(MCID) was proposed to distinguish effects sizes that have a meaningful effect on the patient from those that are simply statistically significant [409]. The initial definition of MCID has been much discussed and refined, so that now there are multiple terms with similar meanings (e.g. minimal clinically important difference, minimally detectable difference), and terms that are similar but have different meanings (e.g. minimally detectable difference versus minimal detectable change) [410]. A full discussion of this concept is beyond the scope of this thesis, and the term MCID will be used for simplicity. However, of relevance is the way MCIDs (and similar measures) are calculated. Broadly, these can be categorised into anchor-based methods (that is comparing the change to an external criterion such as the patient's own perception of the change) and distribution-based methods which rely on the statistical properties of the study's result [411]. This latter can utilise the standard error or the standard deviation, and it has been noted that the MCID of a large variety of health-related quality of life instruments is close to half the standard deviation [412]. It can therefore be seen that accurate measures of the variation of a measurand are required for at least one of the categories of methods of estimating MCID. The much larger estimates of within-individual variation found in this thesis than previously reported suggest the need for a reassessment of the values of MCIDs, and whether MCIDs can be relied upon at all. It is of interest that doctors have been shown to rely less on small changes in HbA1c when making clinical decisions than nurses, and doctors were more aware of the inherent uncertainty in HbA1c results than nurses. Even so, doctors tended to use a cut off of 0.5% (5.5mmol/mol) as a prompt to make therapy changes, which could clearly be due to chance based on the within-individual variation of HbA1c seen in this thesis [333]. For example, a patient with a true mean of 58mmol/mol (7.5%) has an approximately

60% chance of having a measured result further away from the true mean than the MCID.

10.4.1.3 Implications for clinicians

The large within-individual variation of a variety of clinical tests as demonstrated in this thesis suggests a vital need for clinicians to be aware of this problem and how it might affect clinical practice. Although there is no easy solution to improving the accuracy of diagnostic, prognostic and treatment decisions with respect to within-individual variation, some recommendations can be made.

The most important recommendation is to raise awareness among medical practitioners of the issue of within-individual variation. Repeating a test at a different time and averaging the results reduces the impact of within-individual variation, but does not eliminate it. Increased repeat testing may in fact mislead clinicians to assume there is poorer control of a disease than is actually the case because of the higher likelihood of a “chance” high or low value due to within-individual variation [413]. While some doctors seem to be aware of the concept of uncertainty in measurement [333], it is common for physicians to be insensitive to the “reliability of information, or to the degree to which information allows an accurate judgement to be made [414].” It is particularly unclear whether they understand the concept of within-individual variation, or if they do, whether they appreciate its magnitude and ubiquity, not least because of the lack of published data on the within-individual variation of non-laboratory tests. The push towards evidence-based medicine in recent decades, while undoubtedly beneficial in the main, may encourage clinicians to rely more heavily on objective tests rather than their own intuition regarding diagnosis and management. If within-individual variation (among other causes of diagnostic uncertainty) affects the accuracy of the test result without the clinician

being aware of this uncertainty, this may in my opinion lead to errors in diagnosis, prognosis and management, with the adverse patient outcomes that these errors bring. Suggestions for further research into the presentation of uncertainty in clinical data are discussed in section 10.5, Research Recommendations.

It is therefore important that clinicians do not depend too heavily on specific test results when making diagnostic choices, but base their judgement on more global patient assessments such as reported symptoms, physical examinations and other test results.

10.4.1.4 Implications for patients

Consideration needs to be given to the sharing of uncertainty of diagnosis with patients. The past tendency for doctors to behave paternalistically towards patients would mean that they would prefer not to cause the patient anxiety by sharing doubts about diagnosis and treatment. Modern ethics recommend non-paternalistic shared decision-making [415]. A deliberative model has been recommended in which the patient is informed of the results, the results are interpreted for them, and an informed decision is made by the patient after discussion with the clinician [416]. However, uncertainty in diagnosis is a highly complicated concept, with clinicians themselves showing limited understanding of post-test probability [417]. If levels of uncertainty are to be shared with patients, clinicians must be sure they are as accurate as possible. Furthermore, it must be considered whether informing patients of the uncertainty around diagnosis might make informed decisions more difficult, and might lead to suboptimal outcomes. Qualitative research in this area might help inform the ethical decisions as to whether to inform patients about the full extent of medical uncertainty.

10.4.1.5 Implications for reporting results

As noted above, uncertainty in diagnosis is a highly complex subject which is not always fully understood by clinicians. Even those clinicians who do understand concepts such as biological variation, likelihood ratios and positive and negative predictive values cannot be expected to calculate these probabilities in the short space of time available to them in consultations. It would therefore be of benefit if levels of uncertainty were provided automatically to clinicians, either calculated by computerised systems in clinics or provided by laboratories. Currently, limits of uncertainty are only given as a reference range pertaining to the central 95% of a population. Giving clinicians an estimate of uncertainty that incorporates estimates of biological variation and pre-test probability to give more accurate predictive values would allow the clinician to make improved holistic diagnostic and treatment decisions, and reduce their reliance on single test results. Consideration would need to be given as to how this data was presented to the clinician, for example providing ranges of possible results such as 95% confidence intervals rather than point estimates, or a graphical representation with colours or shading demonstrating corresponding to more and less likely results.

10.4.1.6 Implications for creating guidelines

There appears to be a general lack of awareness of the problem of within-individual variation when designing clinical guidelines. For example, as mentioned above, NICE guidelines for diabetes recommend only a single HbA1c test, although suggest that a second test may be “sensible” [328]. The American Gastroenterology Association guidelines for the management of Crohn’s disease suggest a cut-off of CRP of 5mg/L for deciding whether to use endoscopic evaluation of a patient [346].

No mention is made in these guidelines of the within-individual variation of CRP and no recommendation is made for repeat testing of CRP to improve the likelihood that the measured result represents the patient's true mean. NICE guidelines for spirometry recognise that FEV1 levels can show small spontaneous fluctuations, and that over-reliance on a single reversibility test may be misleading unless the change is greater than 0.4 L [354]. However, using the estimates from the data in Chapter eight to simulate a hypothetical patient with a true mean FEV1 of 3.5L (who would have a CV_T of approximately 0.15), the probability of the measured FEV1 being 0.4L higher or lower than 3.5L is 0.42. This suggests the guidelines do not go far enough when advising clinicians to be alert to chance within-individual fluctuations in FEV1 levels. NICE guidelines on depression recommend the use of PHQ-9 or a similar depression symptom measurement instrument to evaluate depression, but make no mention of the possibility of within-individual variation in the test results, possibly because almost no data exists on within-individual variation of depression score measurement instruments [367].

It is therefore vital that the authors of clinical guidelines understand the concept and extent of within-individual variation of the tests they recommend, and advise clinicians on how to mitigate the problems this causes.

Similarly, as discussed above with regards to reporting results, it would be helpful if the authors of reference ranges gave more information about the uncertainty around a measurement representing a patient's true mean. Generally, reference ranges are based on values derived from cross-sectional population studies and do not take account of how individual values may vary over time. The production of reference ranges based on 95% confidence intervals give an impression of a much higher degree of accuracy than is the case. Laboratories and

other researchers who give reference ranges for tests should give some indication of how close an individual test result is likely to be to the true patient mean.

Authors of clinical guidelines and experts making recommendations on diagnostic tests should give consideration to recommendations on the frequency of testing. One conclusion to be drawn from the results of this thesis and other research on biological variation is that more frequent testing should be done. This could involve testing of the healthy population to establish personalised reference ranges, and/or an increased number of tests in diagnosis or monitoring of disease to reduce reliance on a single test that might not represent the true patient mean. If more frequent tests are performed, however it is important that previous tests are taken into account and not just the latest one, to avoid the situation where one test result out of a large number of repeated measurements exceeds a threshold by chance and moves a patient irreversibly into a new categorisation of disease or treatment. It has been shown that repeated tests of HbA1c will inevitably lead to a diagnosis of prediabetes or diabetes if performed frequently enough [405].

However, there are conflicting opinions as to whether more frequent testing reduces the problem of within-individual chance variation or not. One study used statistical modelling of monitoring of repeat testing of cholesterol and blood pressure and concluded that more frequent testing is more likely to produce results reflecting the measurement error rather than the true results [418]. Several studies report on what they consider unnecessary repeat testing, with the potential for patient harm and the addition of significant financial burdens to the healthcare system [419] [420] [421].

Nevertheless, some researchers with an interest in biological variation contradict these arguments. As discussed in Chapter 1.6.2, Fraser [65] considered that many clinical guidelines did not take within-individual variation into account when making recommendations, and recommended that calculations were performed to ascertain how many repeat tests were needed to improve the accuracy of the results. Note that as well as the number of tests that need to be repeated based on the within-individual variation, the timing of repeat testing needs to be carefully considered, taking into account the rapidity with which the measurand varies (e.g. very rapid changes in CRP compared to slow changes in HbA1c), how much the underlying disease varies and how critical it is that the test is highly accurate. Other authors have also noted that repeating and averaging test results reduces the impact of within-individual biological variation [422]. However, it may only be necessary to repeat tests that are near enough to a threshold for decision making (such as making a diagnosis or recommending an intervention) that the measurand's within-individual variation makes it reasonably likely that the true value might be the other side of the threshold. It should also be noted that averaging tests does not take into account material changes in the disease between tests, but not taking previous tests into account runs the risk described above of making erroneous decisions based on outlying results. Ultimately, the decision to repeat a test is a clinical one and is likely to be based on a number of factors which include: cost and patient harm of the test; within-individual variation of the measurand; previous test results; prior probability of the disease or a change in the disease process. Although this decision is ultimately that of the clinician, it is important that much more guidance, in clearly understood terms, are given as to how often to repeat tests, at what interval, and what is the level of uncertainty associated with the results.

10.5 Research Recommendations

The findings in this thesis raise some important questions for interpretation of clinical measurements and their use in clinical decision-making. Firstly, in order for clinicians to make accurate diagnoses, prognoses and treatment recommendations, they must know how likely the measurement they see represents the true status of the patient. In order to know the degree of uncertainty in a measurement, they need to know not only the analytical variation and the variation between individuals, but the variation within-individuals over time. This latter information is often poorly-generalisable, based on poor quality evidence, or completely lacking. It is therefore vital that all clinicians are provided with an estimate of uncertainty that includes the within-individual variation, and therefore this data needs to be described for a much wider range of tests. There is a particular lack of data on within-individual variation for physiological measurements such as pulmonary function testing, and there is no comparable database for these kinds of tests, in contrast to the EFLM test for variation that exists for laboratory measurements. The exception is blood pressure, where the within-individual variation is well-described.

This thesis has highlighted the need for further study into statistical methods required to understand variation. Further research is also required to untangle the effects of long versus short-term variation, and of different influences on variation. For example, do BMI and deprivation have independent effects on variation, or are they proxy measures for diabetic status? Advanced or even new statistical methods may be required to evaluate this. Similarly, more work is needed to account for

confounding by indication, for example clinicians ordering tests more frequently in patients with more severe disease, and selection bias, in which patients tested are at higher risk of the disease for which the test is being performed. Prospective, controlled studies of variability in patients considered at high or low risk of disease could help ascertain the degree to which selection bias plays a role in the results of this study. Studies in which the results of tests are correlated with other markers of the underlying disease might help distinguish increased variation due to pathology from a material change in the disease.

Statistical methods for performing meta-analyses of coefficients of variation are needed to pool data from multiple studies on variation.

It would be useful to develop and validate a risk of bias tool that was suitable to assess risk of bias of variability studies of any type, including different metrics of variation and non-laboratory tests such as PROMs and physiological/functional tests.

How clinicians use within-individual variation to make decisions is uncertain. Personal communications with clinicians suggest there is an awareness of the problem based on clinical experience, but not of the extent of the problem, nor a method to compensate for the problem. Further studies could use clinical vignettes to provide clinicians with data on within-individual variation and then assess how they use this information to make decisions. Alternatively, it might be possible to use databases such as IMRD and CPRD to find out whether clinical decisions that are made are appropriate, for example by looking at a test result and an associated clinical decision, and estimating how likely the clinical decision is to be correct based on the within-individual variation of that test.

I have stated it is important that clinicians be aware of variation and other forms of uncertainty when making measurement-based clinical decisions, but to my knowledge there is no research into methods that would improve this situation. Graphical representations of clinical data have been discussed since the 1960s [423] and more recent studies have shown that the format in which data is presented can affect the interpretation of laboratory data by clinicians [424]. Further research into designing and evaluating the utility of graphical representations of data with associated uncertainty estimates might be helpful.

Given the importance of variation in the calculation of some MCIDs, and the fact that within-individual variation can cause MCIDs to be exceeded by chance variation, the impact of within-individual variation on the calculation and interpretation of MCIDs needs to be investigated further.

Further research is required into the effect of variation on PROMs, and whether PROMs are able to take within-individual variation into account when considering the accuracy of their results, or whether for example other measurement scales are needed in order to incorporate the possibility of within-individual variation. To understand the effect of variation on categorical and integer scales may require new statistical methods.

When performing research into within-individual variation, it is important that demographic factors such as age, sex, deprivation and ethnicity are accounted for. The systematic reviews in this thesis showed a lack of ethnicity data in studies of within-individual variation, which could lead to detrimental health outcomes for Black, Asian and Ethnic Minority groups.

Properly understanding and applying the knowledge of within-individual variation may aid the development of personalised medicine. This involves the use of an individual's own data to develop reference ranges, instead of, as is current practice, utilising reference ranges based on population data. The concept of "personalised laboratory medicine" has recently been described, which takes into account an individual's data such as genomics, clinical chemistry, endocrinology and haematology as well as the within-individual variation of laboratory data [425, 426]. This could of course be extended outside the laboratory to other clinical tests. Smartphones and wearable medical devices provide a large amount of personalised data, which could lead to accurate estimates of personal within-individual variation, and a statistical method has been proposed to analyse the huge longitudinal datasets that this technology can generate [427]. Further research could examine the extent and importance of variation in these personalised datasets.

Within-individual variation is known to be associated with prognosis and outcome in certain conditions. For example, blood pressure variation is associated with outcome in diseases such as cardiovascular disease and dementia (see comparison to other research, above) and HbA1c variation is strongly associated with all-cause mortality in patients with diabetes [428]. The effect of within-individual variation of the measurands in this thesis (and other clinical tests) on prognosis in a variety of disease conditions should be investigated.

Finally, there are a large number of other clinical measurements for which data on within-individual variation is limited or non-existent, particularly non-laboratory measurements such as within-clinic tests and questionnaire-based tests. Further research is needed to evaluate the extent of within-individual variation of a much wider range of clinical tests.

10.6 Conclusion

Systematic reviews of within-individual variation of HbA1c, CRP, spirometry measurements and depression symptom measurement instruments showed that the published data of variation for non-laboratory tests was sparse or absent, and the data that was published tended to be of low quality because of low numbers of participants and/or low numbers of repeat measurements. There was more published data on the variation of HbA1c and CRP, but much of this was also low quality, and there was little published data based on real-world settings, meaning that the generalisability of these studies were low.

Estimates of within-individual variation of these measurands were calculated from large primary care databases, IMRD and CPRD. These studies involved much larger number of participants than previous studies and were based on the real-world data upon which clinical decisions are made. These estimates of variation were much higher than previously published, and it is shown that there is considerable chance of error in clinical decision-making in laboratory tests, non-laboratory tests and patient reported outcome measures if within-individual variation is not taken into account. It was further demonstrated that in general, within-individual variation increases with worsening disease severity.

More data on the size of within-individual variation in real-world settings is needed. Clinicians, designers of clinical trials and authors of clinical guidelines need to take within-individual variation into account when making recommendations.

11 Declarations

The work presented in this thesis is the result of original research carried out by the author Alexander Gerard Gough. During the course of the postgraduate study within the Institute of Applied Health Research, University of Birmingham, the following studies were accepted for publication. Where listed, the secondary authors also advised on study design, data analysis and editing.

Gough A, Sitch A, Ferris E, Marshall T. Within-subject variation of HbA1c: A systematic review and meta-analysis. PLoS One. 2023 Aug 2;18(8):e0289085. doi: 10.1371/journal.pone.0289085.

Gough, A, Marshall, T, Ferris, E & Sitch, A, Within-individual variation of measured depression symptoms: A systematic review and meta-analysis, Journal of Affective Disorders Reports, Volume 14, 2023, 100675, ISSN 2666-9153, <https://doi.org/10.1016/j.jadr.2023.100675>.

Furthermore, a version of Chapter three: Within-individual variation of C-reactive protein and high-sensitivity C-reactive protein: A systematic review and meta-analysis, has been accepted for publication in PLoS One.

I declare that I have no competing interests.

This work was self-funded.

Appendices

Note that appendices are numbered to correspond to the chapter which the data they contain refers to.

12 Appendix 2 HbA1c systematic review appendices

12.1 Appendix 2.1 Search strategies

Search strategies

A. (Medline, Embase, Medline and In-Process, In-Data-Review & Other non-indexed citations)

Sample search strategy, Medline 1946-Nov 2022 and In-Process and other non-indexed citations

1. (CRP or c-reactive protein).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

2. Glycated Hemoglobin A/ or HbA1c.mp.

3. biological varia\$.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

4. within-individual varia\$.mp.

5. within-subject varia\$.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

6. within-individual CV.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

7. Within-subject CV.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

8. Intraindividual CV.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

9. intra-individual varia\$.mp.

10. day-to-day.mp.

11. visit-to-visit.mp.

12. Coefficient of variation.mp.

13. (cvi or cvw or cvb or cvwithin-subject or cvbiological).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol

supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

14. chronic venous insufficiency.mp.

15. vaccine immunology.mp.

16. content validity index.mp.

17. cerebral visual impairment.mp.

18. community virtual ward.mp.

19. cardiac vagal withdrawal.mp.

20. coronary vessel wall.mp.

21. variability independent of the mean.mp.

22. VIM.mp.

23. thalamus.mp. or Thalamus/

24. lactamase.mp. or beta-Lactamases/ or beta-Lactamase Inhibitors/

25. index of individuality.mp.

26. Reference Change Value.mp.

27. rcv.mp.

28. (intraclass or ICC).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

29. Revue canadienne du vieillissement.mp.

30. recoverin.mp. or Recoverin/

31. index of heterogeneity.mp.

32. validity coefficient.mp.

33. reliability parameter.mp.

34. (animal not human).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

35. 1 or 2

36. 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 25 or 26 or 28 or 31 or 32 or 33

37. 14 or 15 or 16 or 17 or 18 or 19 or 20 or 23 or 24 or 29 or 30

38. 35 and 36

39. 38 not 37

40. 39 not 34

12.1.1.1 B. Cochrane central

(CRP) or (HbA1c)

AND

variability or variation or within-individual or within-subject or day-to-day or test-retest

12.1.1.2 C. *Epistemonikos*

(title:((CRP or HbA1c) AND variability) OR abstract:((CRP or HbA1c) AND variability))

D. Open Grey

(CRP or HbA1c) and variability

12.2 Appendix 2.2 Databases searched

Table 12.1 Databases searched

Database	Coverage
Ovid	
Medline	1946 to 11/22
Embase	1974 to 11/22
Medline and In-Process, In-Data-Review & Other non-indexed citations	1946 to 11/22
Cochrane Central	
Epistemonikos	Inception to 11/22
Open Grey	Inception to 11/22
	Inception to 11/22

12.3 Appendix 2.3a Risk of bias template and explanatory notes

Domains assessed were divided into two groups. Group A addressed elements of the research question, which included the name of the instrument, the measurement property (variation measure), and the source of the variation and the patient population. Group B scored risk of bias using 7 questions: were the patients stable?; was the time between measurements appropriate?; were the measurements similar for the repeated measurements except for the source of the variation (in this case time)?; was the measurement administered without knowledge of previous scores in the patient?; were scores assigned without knowledge of previous scores in the patient?; were there any other important flaws in the design or statistical methods?; was a recognised measure of variation calculated with the method of calculation shown? Each of the 7 questions of group B were scored with the following system: 1=very good; 2=adequate; 3=doubtful; 4=inadequate.

Table 12.2 Risk of Bias form, Part A: Elements of Research Question

Element	Instruction	Result
1. Name of the instrument	CRP, hsCRP or HbA1c	
2. Version or way of operationalization	Method of analysis	
3. Construct	Description of what is being measured	
4. Measurement property.	Variation measure(s)	
5. Components that will be repeated	Whole measurement? Parts?	Whole
6. Source(s) of variation varied	Component which is varied across the measurements (time?) Unintentional sources of variation	Time
7. Patient population		

Notes:

4. Reliability study (relative e.g. ICC) or measurement error e.g. SD, CV, and which measurement

6. Any deficiencies that may introduce unintentional sources of variation e.g. training, preparation, collection, transport, test methodology

Table 12.3 Risk of Bias form, Part B: Risk of bias rating

Design requirements	Rating*	Reason/evidence
1. Were patients stable in the time between the repeated measurements on the construct to be measured?		
2. Was the time interval between the repeated measurements appropriate?		
3. Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?		
4. Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?		
5. Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?		
6. Were there any other important flaws in the design or statistical methods of the study?		
7. Was the Coefficient of Variation (CV), ICC, RCV, SD, VIM, variance, ASV, II calculated?		

* Very good/adequate/doubtful/inadequate/NA

Ratings are as follows unless described otherwise below:

Very good: yes; Adequate: reasons to assume standards were met; Doubtful: unclear; Inadequate (no)

1. Completely stable is very good, long-term illness is adequate, short-term/acute illness is doubtful, unstable should have been excluded at screening

2. Any time interval is allowed for the purposes of this review

3. Environmental change may count as doubtful

6. Very good: no; Doubtful: minor methodological flaws; inadequate: yes

7. Very good: the model or formula for the variation measure is described; it matches the reviewer constructed research question and the data;

Adequate: the model or formula is not described or does not optimally match the reviewer constructed research question and evidence provided that no systematic difference has occurred

Doubtful: no knowledge about systematic difference or with evidence provided that systematic difference has occurred

Inadequate: variation calculated based on Cronbach's alpha, or using SD from another population

12.4 Appendix 2.3b Alternative risk of bias template

Quality assessment checklist

Phase 1: State the review question

Patients (setting, intended use of index test, presentation, prior testing):

Phase 2: Risk of bias and applicability judgments

DOMAIN 1: PATIENT SELECTION

Is there a clear description of patient characteristics including healthy/unhealthy?

Yes/No/Unclear

Is there a clear description of setting, locations and relevant dates including follow-up? Yes/No/Unclear

Were the eligibility criteria and methods of selection of patients described?

Yes/No/Unclear

Are all included individuals in steady state, or have data been adequately transformed? Yes/No/Unclear

Were the included patients representative of the population of interest?

Yes/No/Unclear

Did the study avoid inappropriate exclusions? Yes/No/Unclear

Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST(S)

If more than one index test was used, please complete for each test.

Were the number and timing of measurements recorded?

Are the measurand and the measurement procedure documented?

Are preanalytical procedures described and standardised to minimise preanalytical variation?

Could the conduct or interpretation of the index test have introduced bias?

RISK: LOW /HIGH/UNCLEAR

DOMAIN 3: FLOW AND TIMING

☐ Were all patients included in the analysis? Yes/No/Unclear

Could the patient flow have introduced bias? RISK: LOW /HIGH/UNCLEAR

DOMAIN 4: STATISTICAL ANALYSIS

Was the method to determine sample size described? Yes/No/Unclear

Were all statistical methods used described? Yes/No/Unclear

Are confidence limits around CV_I presented? Yes/No/Unclear

Is the number of results excluded following analysis of outliers given, or can they be deduced from the presented data? Yes/No/Unclear

Were the methods to deal with missing data described? Yes/No/Unclear

Are mean concentrations of the measurand presented or can they be extracted?
Yes/No/Unclear

Could the statistical analysis have introduced bias? RISK: LOW/HIGH/UNCLEAR

DOMAIN 5: CONFLICTS OF INTEREST

Was the source of funding noted? Yes/No/Unclear

Were conflicts of interest declared? Yes/No/Unclear

Could conflicts of interest have introduced bias? RISK: LOW/HIGH/UNCLEAR

12.5 Appendix 2.4 Included studies and characteristics

Table 12.4 Included studies and characteristics

Paper	Number of subjects	Age (primary analysis)	% male (primary analysis)
Agard 1997 [429]	442	35	53.17
Barton 2019 [430]	170		0.00
Borsey 1982 [431]	33	52	89.29
Bouchi 2012 [432]	689	65	57.18
Braga 2011 [433]	18		50.00
Cardoso 2018 [434]	654	60.1	38.10
Carlsen 2011 [144]	15		40.00
Ceriello 2017 [435]	4231	67.4	50.90
Chiu 2020 [149]	83	64.5	59.04
Chiu 2017 [436]	27257		46.77
Coskun 2018 [437]	36		52.78
Critchley 2019 [438]	58832	67.7	55.30
Dergez 2014 [439]	20		
Desmeules 2010 [440]	36	14	66.67
Dorajoo 2017 [441]	1170	56.1	52.39
Ercan 2015 [442]	15		
Ercan 2018 [443]	12	43.5	75.00
Foo 2017 [444]	226	62	87.70
Garde 2000 [103]	21		0.00
Godsland 1985 [445]	10		60.00
Gu 2018 [446]	201	65.2	60.20
Gu 2017 [447]	505	69.6	52.08
Hashimoto 2020 [448]	317	66.7	56.78
He 2015 [449]	NR		
Herath 2016 [450]	146		
Hermann 2014 [451]	35891	16.2	52.30
Hietala 2013 [452]	1346	38.7	52.08
Hirakawa 2014 [453]	4399	65.5	42.60
Howey 1989 [454]	8		37.50
Hsu 2012 [455]	821		46.10
Jang 2019 [456]	6462	50	50.00
Jorde 2000 [114]	214	30.4	62.15
Jun 2015 [457]	63	59.5	63.49
Jury 1983 [458]	4		
Kelly 2019 [171]	90	52.8	31.00
Kilpatrick 1998 [459]	12	40	58.33
Kilpatrick 2008 [460]	1441	27	52.80
Kitaoka 2016 [461]	161	62.6	47.50
Lee 2020 [462]	1383	65.6	52.00
Lenters-Westra 2015 [129]	21		
Lenters-Westra 2014 [463]	21		52.38
Li 2017 [464]	15875	69	45.12
Liang 2019 [113]	19	26	47.37
Lin 2013 [465]	3220	57.78	48.91
Low 2017 [466]	1628		57.90
Lui 2020 [150]	83282	71.3	46.20
Luk 2013 [467]	8439	57	47.00
Ma 2012 [468]	881	60	47.90
Mao 2022 [469]	1240	49	52.20
Matsutani 2018 [470]	57	67.2	68.42
Meigs 1996 [471]	639	73.6	35.80

Paper	Number of subjects	Age (primary analysis)	% male (primary analysis)
Mellergard 2020 [141]	158	67.6	65.20
Minami 2020 [472]	118	74.1	58.47
Miyamoto 2008 [473]	18		
Mo 2018 [151]	5278	61	53.00
Moon 2016 [474]	2549	63.49	
Nagl 2014 [475]	37	18.3	
Nazim 2014 [476]	438	8.77	55.02
Nichols 2010 [477]	2417	60.2	48.60
Noguez 2013 [478]	4380		
Parkes 1999 [479]	59		
Parrinello 2015 [480]	61	76	27.00
Pecoraro 1986 [481]	11	62.3	97.56
Penno 2013 [482]	6157	66.4	53.60
Phillipou 1993 [483]	16	48	18.75
Pimentel 2017 [484]	95	43.6	42.11
Prentice 2016 [485]	50,861	65.64	98.00
Ramachandran 2015 [486]	632	51.7	45.40
Raman 2016 [142]	1195	14.7	53.00
Rodriguez-Segade 2012 [487]	2103	59.2	47.70
Rohlfing 2002 [488]	48		
Romero-Aroca 2021 [489]	260	36	51.53
Rosa 2019 [490]	220	29.6	40.00
Rutter 2022 [491]	12860		
Sakai 2020 [492]	301	68	65.78
Sambataro 2017 [493]	39	60	
Savu 2016 [494]	14	62.7	57.14
Schreur 2018 [495]	131	48	46.00
Scott 2018 [496]	9795		
Scott 2019 [497]	506	38	50.00
Segar 2020 [139]	8576	62.7	61.60
Selvin 2007 [498]	685	47.2	49.80
Sheedy 2022 [499]	1102		
Su 2018 [500]	563	56.4	53.10
Sugawara 2012 [140]	619	54.3	69.00
Sun 2019 [501]	930	64.7	54.90
Takahashi 2014 [502]	34662	49	52.00
Takao 2014 [503]	754	54.4	81.80
Takenouchi 2016 [504]	162	62.3	53.60
Teliti 2018 [505]	900	66.96	57.22
Thyagarajan 2016 [506]	57		36.73
Tjaden 2019 [507]	66	14.2	
Trape 2000 [508]	47	61.9	
Tseng 2020 [509]	4216	58.6	49.30
Tsuboi 2017 [510]	168	59.4	61.50
Tsuchiya 2020 [511]	156	67	39.10
Tynjala 2020 [512]	718	46	46.90
Ucar 2013 [513]	29		31.03
Virk 2016 [514]	1706	15.9	
Waden 2009 [515]	2107	36.4	53.20
Walker 2017 [516]	6048	44	
Wan 2016 [138]	91866	62.58	44.20
Watts 1989 [517]	172	31.4	55.81
Whitsel 2012 [518]	100		
Wightman 2018 [519]	5952	44	
Yang 2020 [520]	1705	53.14	57.42
Yang 2018 [521]	681	58.9	60.10
Yang 2015a [522]	31841	60.94	

Paper	Number of subjects	Age (primary analysis)	% male (primary analysis)
Yang 2015b [523]	595	64.88	58.32
Yu 2020 [143]	3254	65	39.34
Zhong 2018 [524]	193	52.17	56.48

Table 12.5 Characteristics of studies (continued)

Paper	Setting code	Health status code	Average number of measurements (primary group)	CV _I or CV _T recorded
Agard 1997 [429]	2	3	16	T
Barton 2019 [430]		2	3	T
Borsey 1982 [431]		2	6	T
Bouchi 2012 [432]	2	2	26	T
Braga 2011 [433]	2	1	5	I
Cardoso 2018 [434]	2	2	12	T
Carlsen 2011 [144]	2	1	10	T
Ceriello 2017 [435]	2	2	5	T
Chiu 2020 [149]	2	2		T
Chiu 2017 [436]		2		T
Coskun 2018 [437]	2	1	4	I
Critchley 2019 [438]	1	2	4	T
Dergez 2014 [439]		1	10	U
Desmeules 2010 [440]	2	5	5	I
Dorajoo 2017 [441]	2	2	11	T
Ercan 2015 [442]		1	7	U
Ercan 2018 [443]	2	1	8	I
Foo 2017 [444]	1	2	4	T
Garde 2000 [103]	2	1	12	T
Godsland 1985 [445]	2	1	11	I
Gu 2018 [446]	2	2	11.7	T
Gu 2017 [447]	2	2	11.3	T
Hashimoto 2020 [448]	2	2	5	T
He 2015 [449]		1	5	U
Herath 2016 [450]		2	2	U
Hermann 2014 [451]	2	3	5	T
Hietala 2013 [452]	2	3	2	T
Hirakawa 2014 [453]	2	2	5	T
Howey 1989 [454]	2	2	6	I
Hsu 2012 [455]	2	2	9	T
Jang 2019 [456]	2	1	5	T
Jorde 2000 [114]	2	3	10.4	T
Jun 2015 [457]	2	2	6	T
Jury 1983 [458]	2	2	18	T
Kelly 2019 [171]	1	1	2	U
Kilpatrick 1998 [459]	2	1	10	I
Kilpatrick 2008 [460]	2	3		T
Kitaoka 2016 [461]		2	12	T
Lee 2020 [462]	2	2	4	T
Lenters-Westra 2015 [129]	2	1	4	U
Lenters-Westra 2014 [463]	2	1	4	I
Li 2017 [464]	1	2	3.46	T
Liang 2019 [113]	2	1	5	I
Lin 2013 [465]	2	2		T
Low 2017 [466]	2	2	8	T
Lui 2020 [150]	2	2	7.1	T
Luk 2013 [467]	2	2	10	T

Paper	Setting code	Health status code	Average number of measurements (primary group)	CV _I or CV _T recorded
Ma 2012 [147]	2	2	12	T
Matsutani 2018 [470]	2	2	8	T
Mao 2022 [469]	1	3		T
Meigs 1996 [471]	1	1	2	T
Mellergard 2020 [141]	1	2	5	T
Minami 2020 [472]	2	2		T
Miyamoto 2008 [473]	2	5	9.1	T
Mo 2018 [151]	2	2	4	T
Moon 2016 [474]		2		U
Nagl 2014 [475]		3		U
Nazim 2014 [476]	2	3	36.8	T
Nichols 2010 [477]	2	2	12.4	T
Noguez 2013 [478]				I
Parkes 1999 [479]	2	4	2	T
Parrinello 2015 [480]	1	1	2	U
Pecoraro 1986 [481]	2	2	3	T
Penno 2013 [482]	2	2	4.52	T
Phillipou 1993 [483]	2	4	4.1	B
Pimentel 2017 [484]	2	5	3	I
Prentice 2016 [485]	1	2	4	T
Ramachandran 2015 [486]	1	2	6.3	T
Raman 2016 [142]	2	3	14	T
Rodriguez-Segade 2012 [487]	2	2	10	T
Rohlfing 2002 [488]		1	12	I
Romero-Aroca 2021 [489]	2	3	8	T
Rosa 2019 [490]	2	3	17	U
Rutter 2022 [491]	2	1	2	T
Sakai 2020 [492]	2	2	5	T
Sambataro 2017 [493]		2		U
Savu 2016 [494]	2	2	3	T
Schreur 2018 [495]	2	3	5	T
Scott 2018 [496]		2	6	T
Scott 2019 [497]	2	3	8	T
Segar 2020 [139]	2	2	8	T
Selvin 2007 [498]	1	1	2	I
Sheedy 2022 [499]	2	4		T
Su 2018 [500]	2	2	4	T
Sugawara 2012 [140]	2	2	11	T
Sun 2019 [501]		2	5	T
Takahashi 2014 [502]	2	1	5	T
Takao 2014 [503]	2	2	4	T
Takenouchi 2016 [504]	2	2	6	T
Teliti 2018 [505]	2	2	5	T
Thyagarajan 2016 [506]	1	1	2	I
Tjaden 2019 [507]	2	2	2	T
Trape 2000 [508]	2	2	4.2	I
Tseng 2020 [509]	2	2	3	T
Tsuboi 2017 [510]	2	2	12	T
Tsuchiya 2020 [511]		2	13	T
Tynjala 2020 [512]	1	3	16	T
Ucar 2013 [513]		1	4	T
Virk 2016 [514]	2	3	22	T
Waden 2009 [515]	2	3	13	T
Walker 2017 [516]	1	3	4	T
Wan 2016 [138]	1	2	6.47	T
Watts 1989 [517]	2	3	8	T

Paper	Setting code	Health status code	Average number of measurements (primary group)	CV _I or CV _T recorded
Whitsel 2012 [518]	1	1	2	U
Wightman 2018 [519]	1	3	4	T
Yang 2020 [520]	2	2		T
Yang 2018 [521]		2	16.9	T
Yang 2015a [522]	1	2	3	T
Yang 2015b [523]	2	2	4	T
Yu 2020 [143]	1	1		T
Zhong 2018 [524]	1	4	12.82	T

Setting code: 1=primary or community;2=tertiary/secondary/laboratory

Health status code: 1=healthy 2= type 2 diabetes 3= type 1 diabetes 4= both type 1 diabetes and type 2 diabetes 5=other disease

CV_I or CV_T recorded code: I= CV_I; T=CV_T; U= Uncertain

12.6 Appendix 2.5 Risk of bias estimates for each paper

Table 12.6 Risk of bias estimates for each included paper, with questions 4 and 5 excluded.

Paper	Risk of Bias
Agard 1997 [429]	2
Barton 2019 [430]	3
Borsey 1982 [431]	1
Bouchi 2012 [432]	1
Braga 2011 [433]	1
Cardoso 2018 [434]	1
Carlsen 2011 [144]	1
Ceriello 2017 [435]	3
Chiu 2020 [149]	2
Chiu 2017 [436]	2
Coskun 2018 [437]	1
Critchley 2019 [438]	2
Dergez 2014 [439]	2
Desmeules 2010 [440]	1
Dorajoo 2017 [441]	1
Ercan 2015 [442]	3
Ercan 2018 [443]	1
Foo 2017 [444]	2
Garde 2000 [103]	1
Godsland 1985 [445]	1
Gu 2018 [446]	3
Gu 2017 [447]	2
Hashimoto 2020 [448]	2
He 2015 [449]	2
Herath 2016 [450]	3
Hermann 2014 [451]	2
Hietala 2013 [452]	2
Hirakawa 2014 [453]	1
Howey 1989 [454]	2
Hsu 2012 [455]	2
Jang 2019 [456]	1
Jorde 2000 [114]	1
Jun 2015 [457]	1
Jury 1983 [458]	3
Kelly 2019 [171]	2
Kilpatrick 1998 [459]	1
Kilpatrick 2008 [460]	2
Kitaoka 2016 [461]	2
Lee 2020 [462]	2
Lenters-Westra 2015 [129]	2
Lenters-Westra 2014 [463]	3
Li 2017 [464]	3
Liang 2019 [113]	1
Lin 2013 [465]	2
Low 2017 [466]	2
Lui 2020 [150]	2
Luk 2013 [467]	3
Ma 2012 [147]	2
Mao 2022 [469]	2
Matsutani 2018 [470]	3
Meigs 1996 [471]	2

Paper	Risk of Bias
Mellergard 2020 [141]	3
Minami 2020 [472]	2
Miyamoto 2008 [473]	3
Mo 2018 [151]	3
Moon 2016 [474]	3
Nagl 2014 [475]	3
Nazim 2014 [476]	2
Nichols 2010 [477]	2
Noguez 2013 [478]	3
Parkes 1999 [479]	3
Parrinello 2015 [480]	3
Pecoraro 1986 [481]	2
Penno 2013 [482]	2
Phillipou 1993 [483]	3
Pimentel 2017 [484]	2
Prentice 2016 [485]	3
Ramachandran 2015 [486]	3
Raman 2016 [142]	2
Rodriguez-Segade 2012 [487]	2
Rohlfing 2002 [488]	3
Romero-Aroca 2021 [489]	2
Rosa 2019 [490]	3
Rutter 2022 [491]	1
Sakai 2020 [492]	2
Sambataro 2017 [493]	3
Savu 2016 [494]	3
Schreur 2018 [495]	2
Scott 2018 [496]	2
Scott 2019 [497]	2
Segar 2020 [139]	2
Selvin 2007 [498]	1
Sheedy 2022 [499]	2
Su 2018 [500]	2
Sugawara 2012 [140]	2
Sun 2019 [501]	2
Takahashi 2014 [502]	3
Takao 2014 [503]	2
Takenouchi 2016 [504]	3
Teliti 2018 [505]	3
Thyagarajan 2016 [506]	2
Tjaden 2019 [507]	3
Trape 2000 [508]	1
Tseng 2020 [509]	3
Tsuboi 2017 [510]	2
Tsuchiya 2020 [511]	3
Tynjala 2020 [512]	2
Ucar 2013 [513]	2
Virk 2016 [514]	2
Waden 2009 [515]	2
Walker 2017 [516]	3
Wan 2016 [138]	2
Watts 1989 [517]	3
Whitsel 2012 [518]	3
Wightman 2018 [519]	3
Yang 2020 [520]	3
Yang 2018 [521]	2
Yang 2015a [522]	2

Paper	Risk of Bias
Yang 2015b [523]	2
Yu 2020 [143]	1
Zhong 2018 [524]	3

Highest score presented. 1=very good; 2=adequate; 3=doubtful; 4=inadequate.

12.7 Appendix 2.6 Statistical calculations for ICC meta-analysis

$Z = 0.5 * \ln((1 + ICC) / (1 - ICC))$. Standard error was calculated using the formula

$SE = \sqrt{1/(N - 3)}$ Z scores were then back-transformed

12.7.1 Stata code for ICC:

```
sort Totalsubjectsinprimaryanalys
```

```
gen z = 0.5 * ln((1 + ICCprimarygroup ) / (1 - ICCprimarygroup ))
```

```
gen sez = sqrt(1/( Totalsubjectsinprimaryanalys - 3))
```

```
meta set z sez, random(dl) studylabel( Shorttitle ) studysize(
Totalsubjectsinprimaryanalys )
```

```
meta forestplot
```

```
gen iccbk = (exp(2*z) - 1)/(exp(2*z) + 1)
```

```
gen cilicc = (exp(2* _meta_cil ) - 1)/(exp(2* _meta_cil ) + 1)
```

```
gen ciuicc = (exp(2* _meta_ciu ) - 1)/(exp(2* _meta_ciu ) + 1)
```

```
metan iccbk cilicc ciuicc if Unitusedforprimaryanalysis ==1 , random by(
```

```
Healthstatusshortname ) label( namevar=Shorttitle) rcols (
```

```
Totalsubjectsinprimaryanalys ) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) xtitle ("ICC")
```

12.7.2 Stata code for CV

```
sort Totalsubjectsinprimaryanalys
```

```
label variable PrimarymeasureofCVIaverage "CVI"
```


rename PrimarymeasureofCVIaverage CVI

label variable Healthstatusshortname "Health status"

label variable Shorttitle "Study title"

label variable Averagenumberofmeasurements "Number of measurements"

label variable Totalsubjectsinprimaryanalys "Number of subjects"

generate cil = CVI - CVI * 0.0001

generate ciu = CVI + CVI * 0.0001

metan CVI cilc ciu if Unitusedforprimaryanalysis ==1 , randomi by(
Healthstatusshortname) label(namevar=Shorttitle) rcols (CVI
Averagenumberofmeasurement Totalsubjectsinprimaryanalys) nosubgroup nostats
xlabel (0, 0.05, 0.1, 0.15, 0.2) nooverall nowt nobox xtitle (CV)

metan CVI cilc ciuc if (Unitusedforprimaryanalysis ==1 & Healthstatuscode ==1),
randomi label(namevar=Shorttitle) rcols (CVI Averagenumberofmeasurement
Totalsubjectsinprimaryanalys) nosubgroup nostats xlabel (0, 0.02, 0.04, 0.06, 0.08,
0.10) nooverall nowt nobox xtitle (CV)

```

metan CVI cil ciu if Healthstatuscode ==2 , randomi by( Unitusedforprimaryanalysis )
label( namevar=Shorttitle) rcols ( CVI Averagenumberofmeasurement
Totalsubjectsinprimaryanalys ) nosubgroup nostats xlabel (0, 0.02, 0.04, 0.06, 0.08)
nooverall nowt nobox xtitle (CV)

```

12.8 Appendix 2.7 Forest plots of subgroup and sensitivity analyses

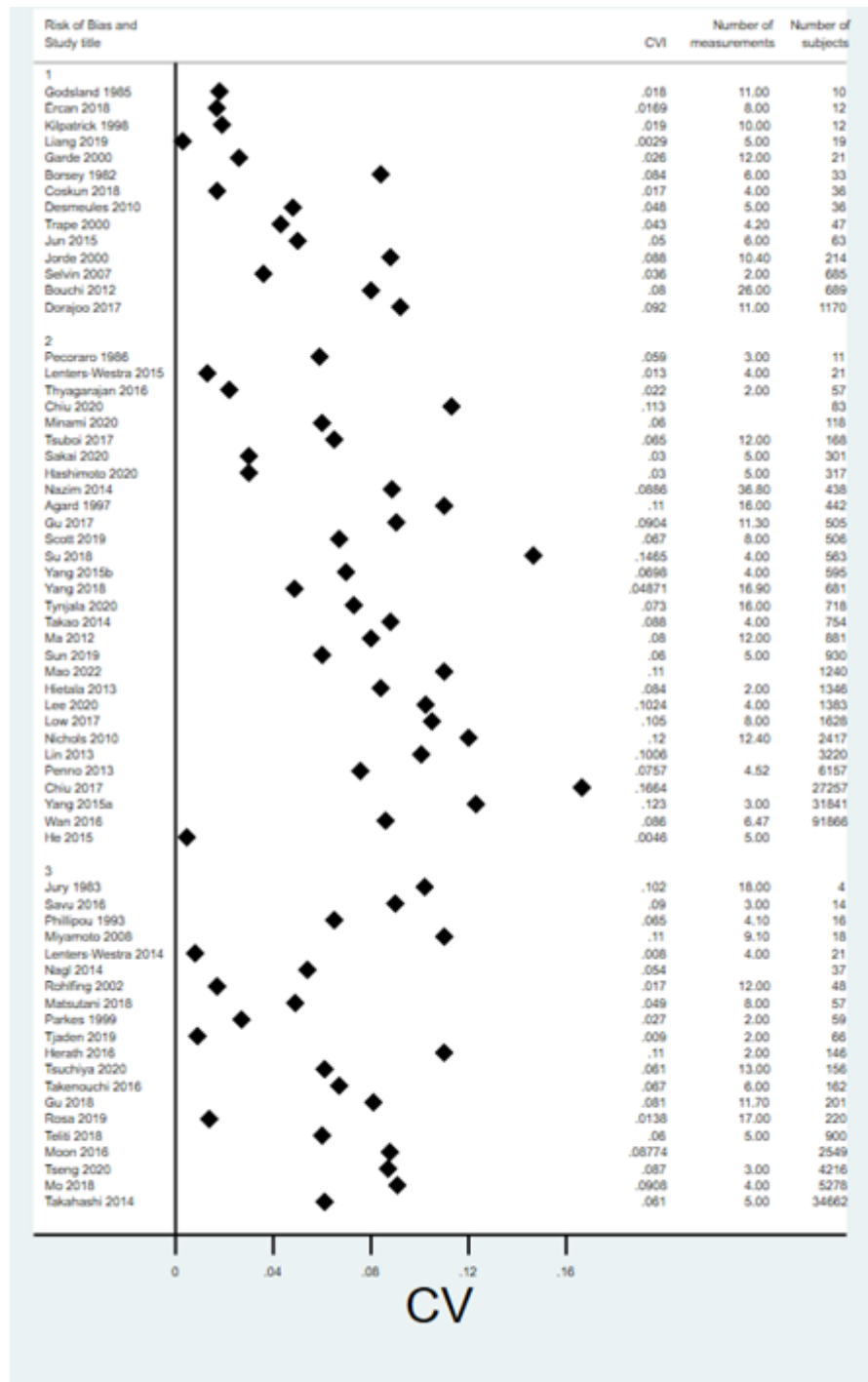


Figure 12.1 Forest plot of CV for studies reporting units of %, split by risk of bias

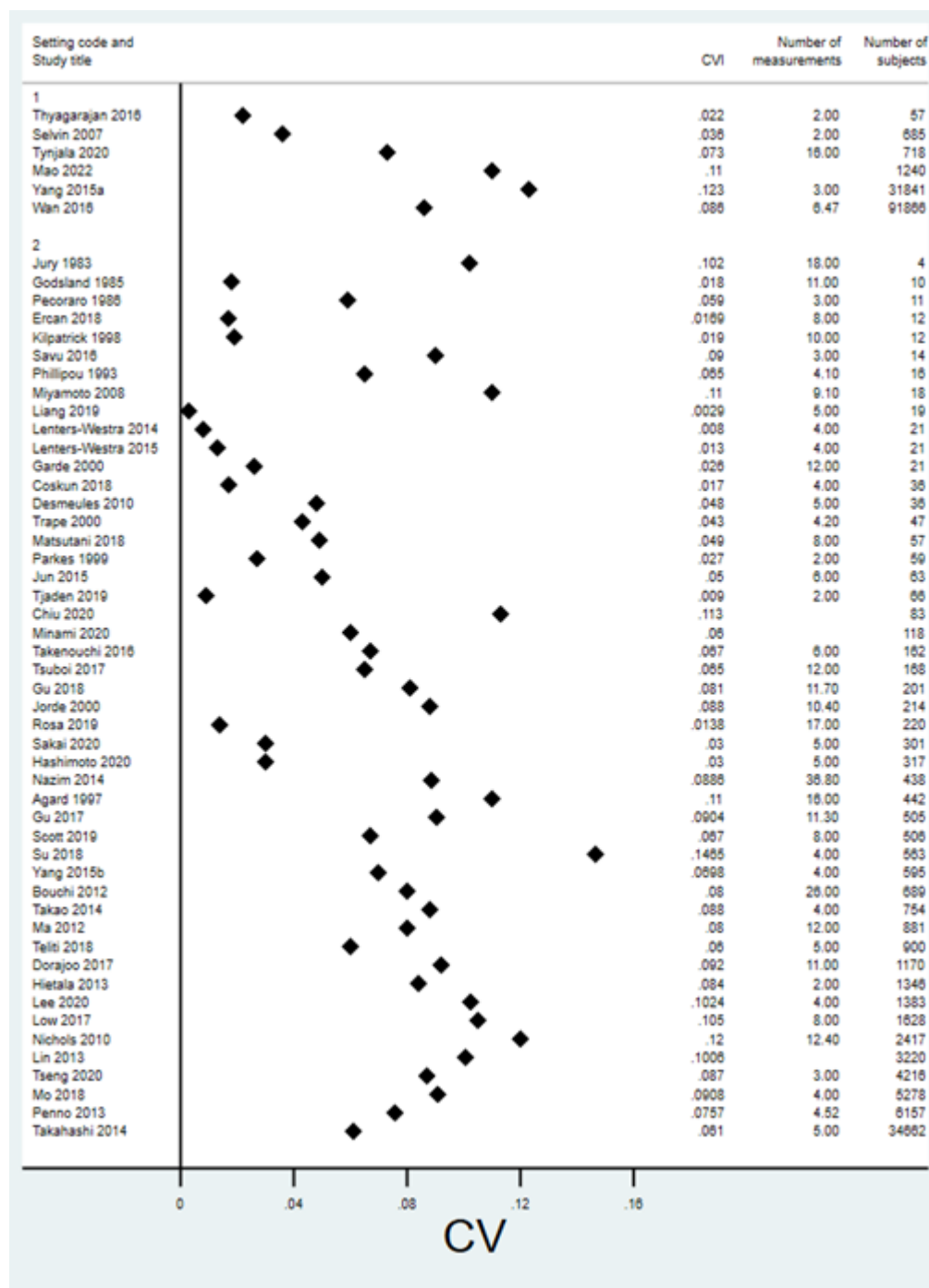


Figure 12.2 Forest plot of CV for studies reporting units of %, split by setting.

1=primary care or community, 2=secondary or tertiary care

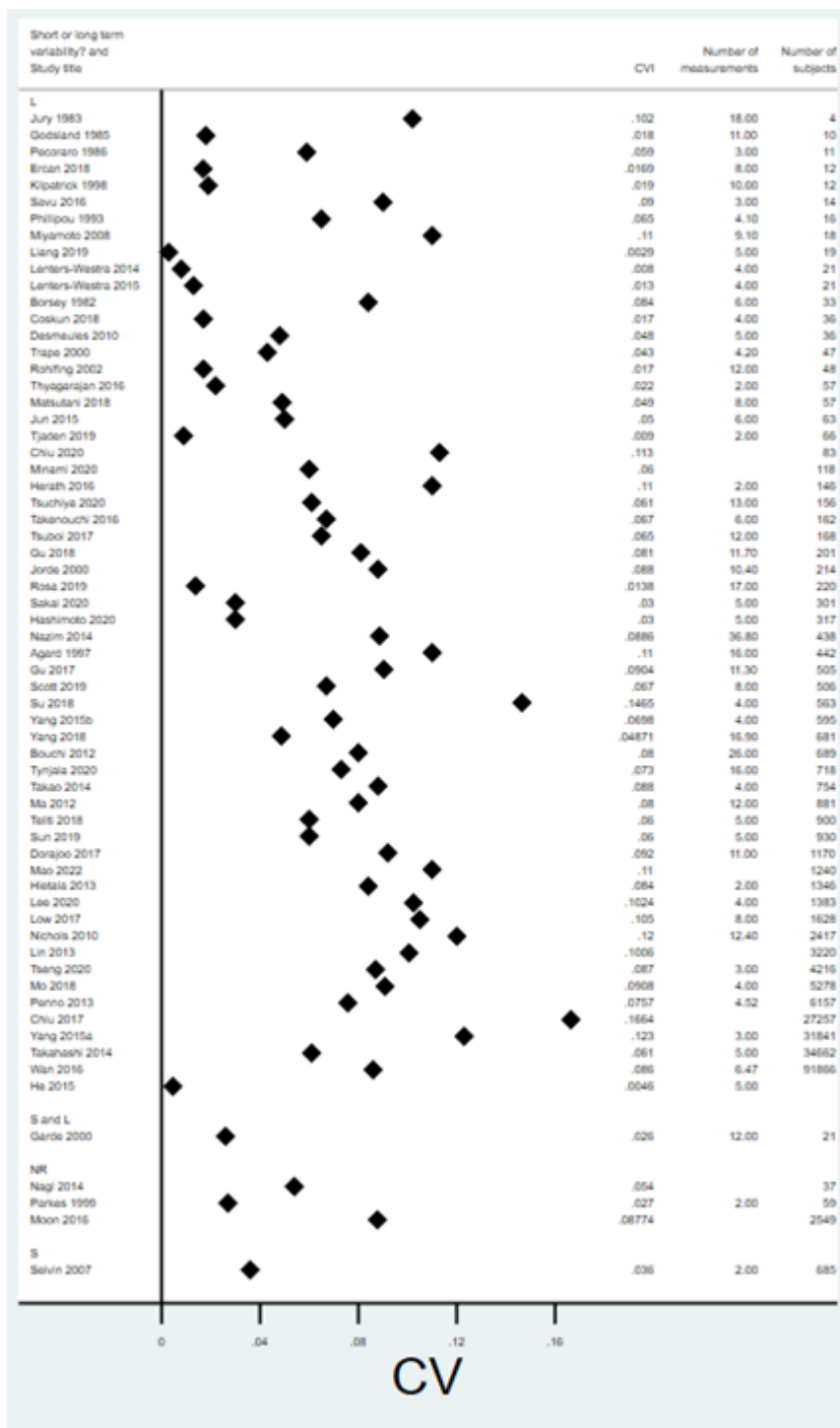


Figure 12.3 Forest plot of CV for studies reporting units of %, split by short (S) or long (L) term measurements

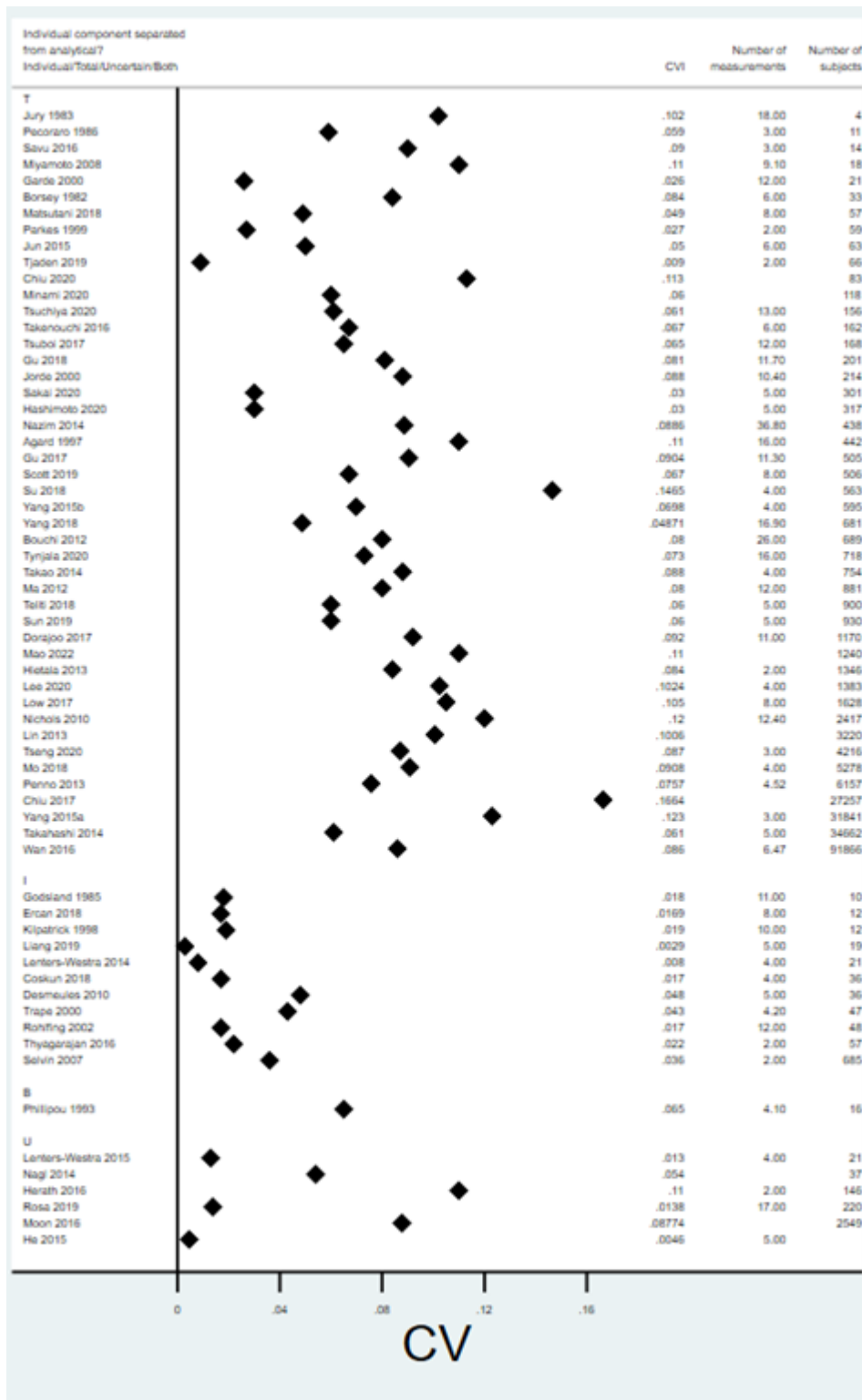


Figure 12.4 Forest plot of CV for studies reporting units of %, split by whether CVI (I) or CVT (T) was reported or whether this was uncertain (U)

13 Appendix 3 CRP systematic review appendices

13.1 Appendix 3.1 Statistical calculations for ICC meta-analysis

$Z = 0.5 * \ln((1 + ICC) / (1 - ICC))$. Standard error was calculated using the formula

$SE = \sqrt{1/(N - 3)}$ Z scores were then back-transformed

Stata code for ICC:

```
gen z = 0.5 * ln((1 + PrimarymeasureofICCseentoe ) / (1 -  
PrimarymeasureofICCseentoe ))
```

```
gen sez = sqrt(1/( Totalsubjectsinprimaryanalys - 3))
```

```
meta set z sez, random(dl) studylabel( Shorttitle ) studysize(  
Totalsubjectsinprimaryanalys )
```

```
meta forestplot
```

```
gen iccbk = (exp(2*z) - 1)/(exp(2*z) + 1)
```

```
gen cil = (exp(2* _meta_cil ) - 1)/(exp(2* _meta_cil ) + 1)
```

```
gen ciu = (exp(2* _meta_ciu ) - 1)/(exp(2* _meta_ciu ) + 1)
```

```
metan iccbk cil ciu if CRPhsCRP==1 , by ( Healthstatusname ) random label(  
namevar=Shorttitle) rcols ( Totalsubjectsinprimaryanalys
```

```
Averagenumberofmeasurementso ) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) xtitle ("ICC")
```

metan iccbk cil ciu if CRPhsCRP == 1, random by (Settingcodetext) label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)

metan iccbk cil ciu if CRPhsCRP == 2, random by (Settingcodetext) label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)

metan iccbk cil ciu if CRPhsCRP == 1, random by (Healthstatusname) label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)

metan iccbk cil ciu if CRPhsCRP == 2, random by (Healthstatusname) label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)

metan iccbk cil ciu if (Totalsubjectsprimaryanalys >100), random by (HbA1cCRPhsCRPormultiple) label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)

metan iccbk cil ciu if (ROBHighest <3), random by (HbA1cCRPhsCRPormultiple) label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)

metan iccbk cil ciu if (CRPhsCRP ==1 & Shortorlongtermvariability == "L"), random label (namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (Totalsubjectsprimaryanalys) xtitle (ICC)


```
metan iccbk cil ciu if ( CRPhsCRP ==2 & Shortorlongtermvariability == "L" ), random
label ( namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6, 0.8, 1.0) rcols (
Totalsubjectsinprimaryanalys ) xtitle (ICC)
```

```
metan iccbk cil ciu if ( CRPhsCRP ==1 ), random by (
Unitofmeasurementofmeasurand ) label ( namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6,
0.8, 1.0) rcols ( Totalsubjectsinprimaryanalys ) xtitle (ICC)
```

```
metan iccbk cil ciu if ( CRPhsCRP ==2 ), random by (
Unitofmeasurementofmeasurand ) label ( namevar=Shorttitle) xlabel (0, 0.2, 0.4, 0.6,
0.8, 1.0) rcols ( Totalsubjectsinprimaryanalys ) xtitle (ICC)
```

13.2 Appendix 3.2 Study characteristics

Table 13.1 Study characteristics detail

Paper	CRP or hsCRP	Number of subjects	Age (primary analysis)	% male (primary analysis)
Alexander 2013 [525]	hsCRP	15		40.00
Bailey 2014 [526]	CRP	27		45.00
Bay-Jensen 2017 [527]	hsCRP	490		
Becker 2020 [528]	hsCRP	11	18	100
Block 2006 [529]	CRP	206	45.4	
Blum 2005 [530]	hsCRP	15		0.00
Boenisch 2002 [531]	CRP	10	67	10.00
Bogaty 2013 [532]	hsCRP	100	65.6	100.00
Bogaty 2005 [533]	hsCRP	159		
Browning 2004 [534]	CRP	15		0.00
Carobene 2019 [535]	CRP	87		42.53
Cho 2005 [536]	CRP	11	30	0.00
Christmas 2018 [537]	hsCRP	50	26	72.00
Clark 1993 [538]	CRP	19		0.47
Corte 2020 [539]	CRP	11		0.27
De Maat 1996 [540]	CRP	20	31	
DeGoma 2012 [541]	CRP	255	59	60.00
D'Eril 2001 [542]	CRP	22		50.00
Engelberger 2015 [543]	hsCRP	80	57	55.00
Franzini 1993 [544]	CRP	10		
Gasco 2010 [545]	CRP	16	52.7	50.00
Glynn 2009 [546]	hsCRP	8901		62.09
Guy 2017 [547]	CRP	12	24.3	100.00
Hardikar 2014 [548]	hsCRP	329	61	82.20
Hosogaya 1999 [549]	CRP	22		0.73
Itoh 2012 [550]	hsCRP	4		100.00
Jackson 2015 [551]	CRP	3410	65	44.10
Jain 2020 [552]	hsCRP	37	22.6	45.21
Jung 2019 [553]	CRP	20		
Kelly 2019 [171]	CRP	90	52.8	31.00
Koenig 2003 [554]	CRP	696		100.00
Kolsum 2009 [555]	CRP	58	63.78	68.97
Laclair 2008 [556]	hsCRP	32	50	100.00
Lee 2007 [557]	hsCRP	48	54.8	10.00
Macy 1997 [558]	CRP	26		38.46
Malaponte 2007 [46]	CRP	18	72.86	55.81
Meyer 2011 [559]	CRP	467	24.9	100.00
Nasermoaddeh 2006 [560]	hsCRP	1679		0.54
Navarro 2012 [561]	hsCRP	62	30.4	50.00
Nunes 2010 [47]	CRP	56	18	100.00
Ockene 2001 [562]	hsCRP	113	49	56.64
Park 2022 [563]	CRP	89	70.5	54
Platz 2010 [564]	hsCRP	50	64.9	100.00
Qi 2016 [565]	hsCRP	40	46	52.50
Riese 2002 [566]	CRP	224	45.2	41.07
Rubin 2018 [567]	CRP	130	43.9	0.00
Rudez 2009 [568]	CRP	40	41	35.00
Rutter 2022 [491]	CRP	16549		
Sakkinen 1999 [569]	CRP	26	27	38.46
Sennels 2007 [570]	hsCRP	38	43	61.00

Paper	CRP or hsCRP	Number of subjects	Age (primary analysis)	% male (primary analysis)
Shaw 2014 [280]	CRP	NR		
Sjoberg 2014 [571]	CRP	188	66	5.00
Thyagarajan 2016 [506]	hsCRP	50		63.27
Thompson 2022 [572]	CRP	272	62.7	37.8
Todd 2013 [573]	hsCRP	25	36	82.00
Tsirpanlis 2004 [574]	hsCRP	29	63.96	51.72
van den Berg 2020 [575]	hsCRP	98	62.8	77.00
Waschki 2010 [576]	hsCRP	136	64	
White 2017 [277]	CRP	249	29.8	46.00
Wu 2012 [577]	hsCRP	56218	51.8	78.52

Table 13.2 Study characteristics detail (continued) Characteristics of studies included.

Paper	Setting code	Health status code	Average number of measurements (primary group)	Variation measure
Alexander 2013	2	1	18.00	I
Bailey 2014	1	1	4.00	I
Bay-Jensen 2017		2	2.00	T
Becker 2020	2	1	18.00	T
Block 2006	2	1	2.00	T
Blum 2005		1	7.00	T
Boenisch 2002		4	12.00	
Bogaty 2013	2	3	15.00	T
Bogaty 2005	2	3	3.00	T
Browning 2004	2	1	3.00	T
Carobene 2019	2	1	10.00	I
Cho 2005		5	10.00	I
Christmas 2018	2	1	2.00	T
Clark 1993	2	1	10.00	I
Corte 2020	2	1	5.00	T
De Maat 1996	2	1	9.00	I
DeGoma 2012	1	1	2.00	T
D'Eiril 2001	2	1	5.00	I
Engelberger 2015	2	1	4.00	T
Franzini 1993	2	1	8.00	I
Gasco 2010		4	12.00	I
Glynn 2009		1	6.00	T
Guy 2017	2	1	2.00	T
Hardikar 2014	2	5	2.00	T
Hosogaya 1999	2	1	12.00	U
Itoh 2012	2	1	6.00	T
Jackson 2015	1	1	2.00	T
Jain 2020	2	2	2.00	T
Jung 2019	2	2	3.00	T
Kelly 2019	1	1	2.00	T
Koenig 2003	1	1	2.00	I
Kolsum 2009	1	2	2.00	T
Laclair 2008	2	4	4.00	I
Lee 2007	1	1	4.00	T
Macy 1997	2	1	8.00	I
Malaponte 2007	2	4	8.00	T
Meyer 2011	2	1	4.00	T
Nasermoaddeh 2006	1	1	2.00	T
Navarro 2012	2	1	4.00	T
Nunes 2010	2	1	4.00	I
Ockene 2001	2	1	5.00	T
Platz 2010	1	1	3.00	T
Park 2022	2	2	2.00	T
Qi 2016	2	1	5.00	I
Riese 2002	1	1	2.50	I
Rubin 2018	1	1	2.5	U
Rudez 2009	2	1	13.00	I
Rutter 2022	2	1	2.00	T
Sakkinen 1999	2	1	9.00	I
Sennels 2007	1	1	6.00	I
Shaw 2014		2		T
Sjoberg 2014	2	4	2.00	T
Thompson 2022	2	5	3.00	T
Thyagarajan 2016	1	1	2.00	I

Paper	Setting code	Health status code	Average number of measurements (primary group)	Variation measure
Todd 2013	2	1	6.00	I
Tsirpanlis 2004	2	4	16.00	I
van den Berg 2020	2	3	4.55	I
Waschki 2010	2	2	2.00	U
White 2017	2	2	2.50	T
Wu 2012	1	1	2.00	T

Setting code: 1=primary or community;2=tertiary/secondary/laboratory. Health status code: 1=Healthy; 2=respiratory disease 3=heart disease 4=haemodialysis 5=other. Variation measure refers to whether the total variation ($CV_I + CV_A$), T, or the individual variation (CV_I), I, was reported, or whether it was uncertain, U.

13.3 Appendix 3.3 Risk of bias scores

Table 13.3 Risk of bias scores Maximum risk of bias score for each study with questions B4 and B5 excluded, and lowest BIVAC score (for studies that report a CV)

Paper	COSMIN Risk of bias score	BIVAC risk of bias score
Alexander 2013	1	C
Bailey 2014	1	C
Bay-Jensen 2017	1	C
Becker 2020	1	
Block 2006	1	
Blum 2005	3	
Boenisch 2002	1	C
Bogaty 2013	1	
Bogaty 2005	2	
Browning 2004	1	
Carobene 2019	1	A
Cho 2005	1	
Christmas 2018	1	C
Clark 1993	1	C
Corte 2020	1	A
De Maat 1996	2	
DeGoma 2012	1	
D'Eril 2001	1	C
Engelberger 2015	1	C
Franzini 1993	1	C
Gasco 2010	2	C
Glynn 2009	1	
Guy 2017	1	C
Hardikar 2014	1	
Hosogaya 1999	1	C
Itoh 2012	1	C
Jackson 2015	2	
Jain 2020	1	
Jung 2019	3	D
Kelly 2019	2	
Koenig 2003	1	
Kolsum 2009	3	
Laclair 2008	1	
Lee 2007	2	
Macy 1997	1	C
Malaponte 2007	1	C
Meyer 2011	2	C
Nasermoaddeh 2006	2	
Navarro 2012	1	
Nunes 2010	1	C
Ockene 2001	1	
Park 2022	2	
Platz 2010	1	
Qi 2016	1	C
Riese 2002	2	C
Rubin 2018	3	
Rudez 2009	1	C
Rutter 2022	1	
Sakkinen 1999	1	C
Sennels 2007	1	C

Paper	COSMIN Risk of bias score	BIVAC risk of bias score
Shaw 2014	3	
Sjoberg 2014	2	
Thompson 2022	2	
Thyagarajan 2016	2	C
Todd 2013	1	C
Tsirpanlis 2004	2	
van den Berg 2020	1	C
Waschki 2010	3	
White 2017	3	C
Wu 2012	2	

13.4 Appendix 3.4 Forest plots of subgroup analyses

13.4.1 CRP CVs

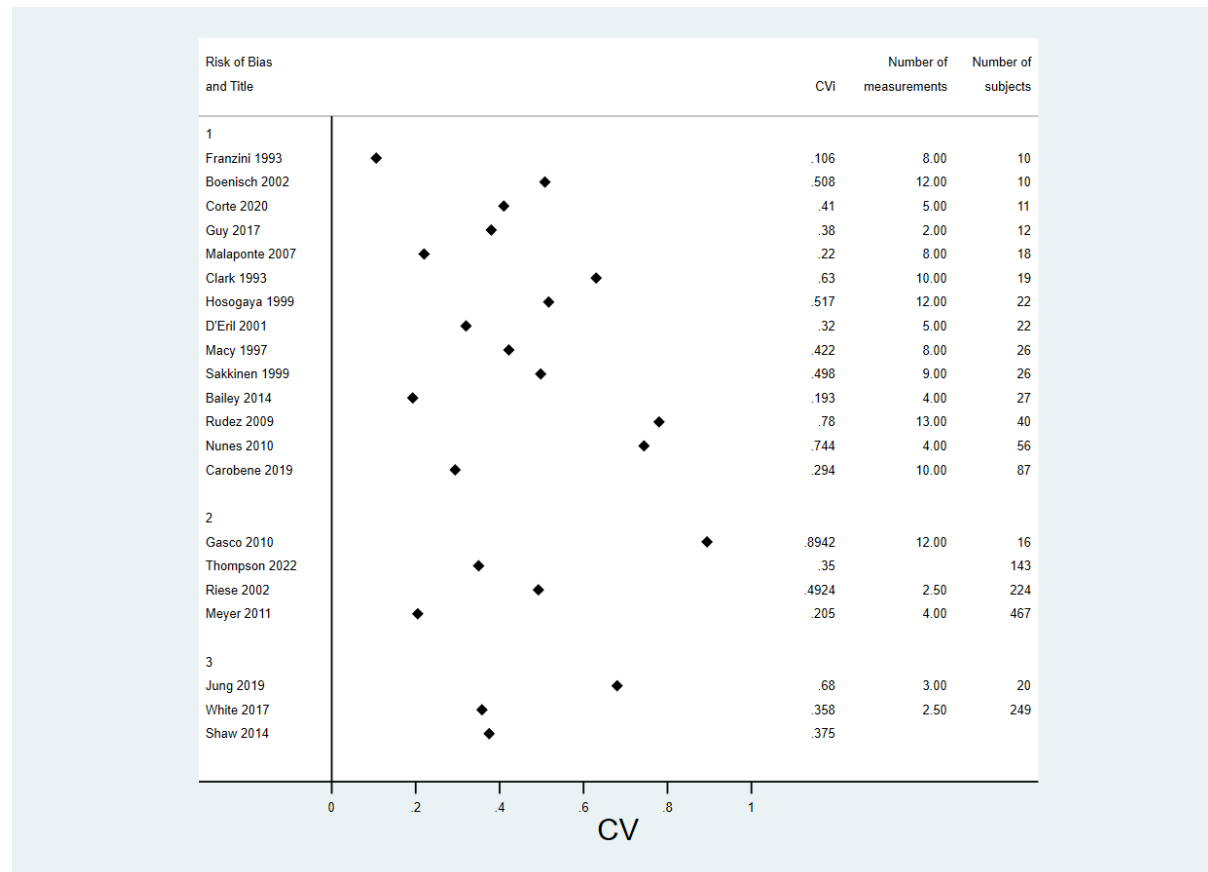


Figure 13.1 Forest plot of all studies that reported a CV for CRP, split by Risk of Bias

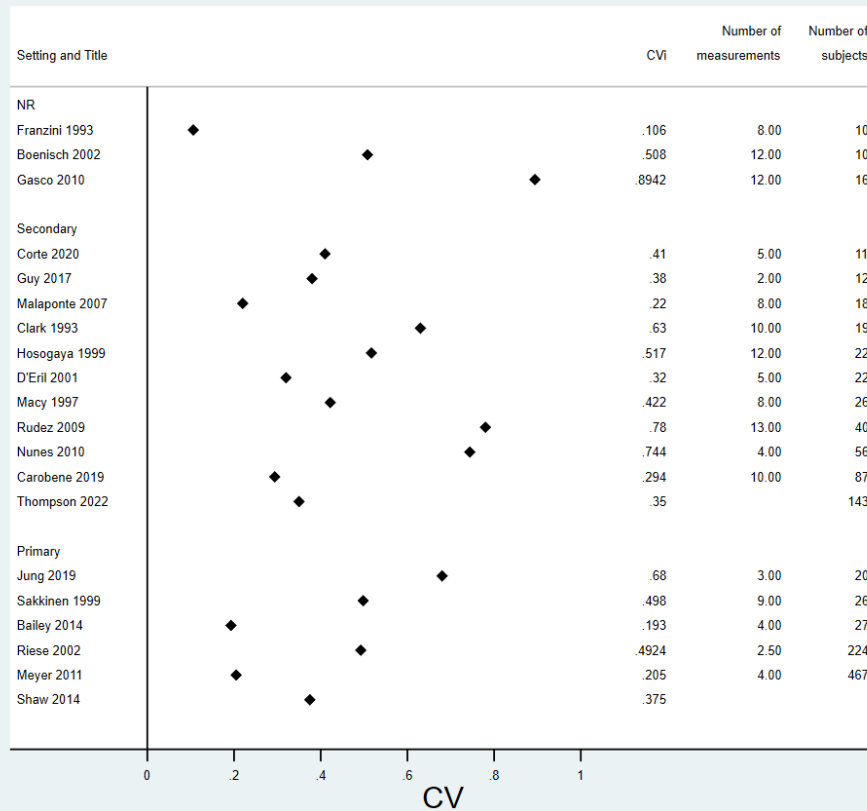


Figure 13.2 Forest plot of all studies that reported a CV for CRP, split by setting

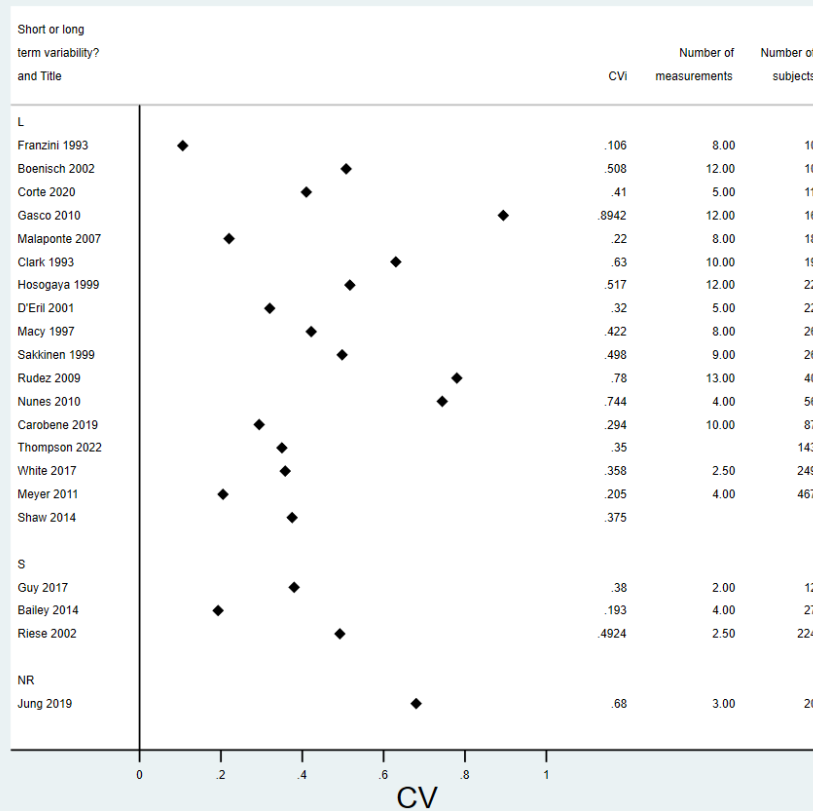


Figure 13.3 Forest plot of all studies that reported a CV for CRP, split by long (L) or short (S) term measurement of variation

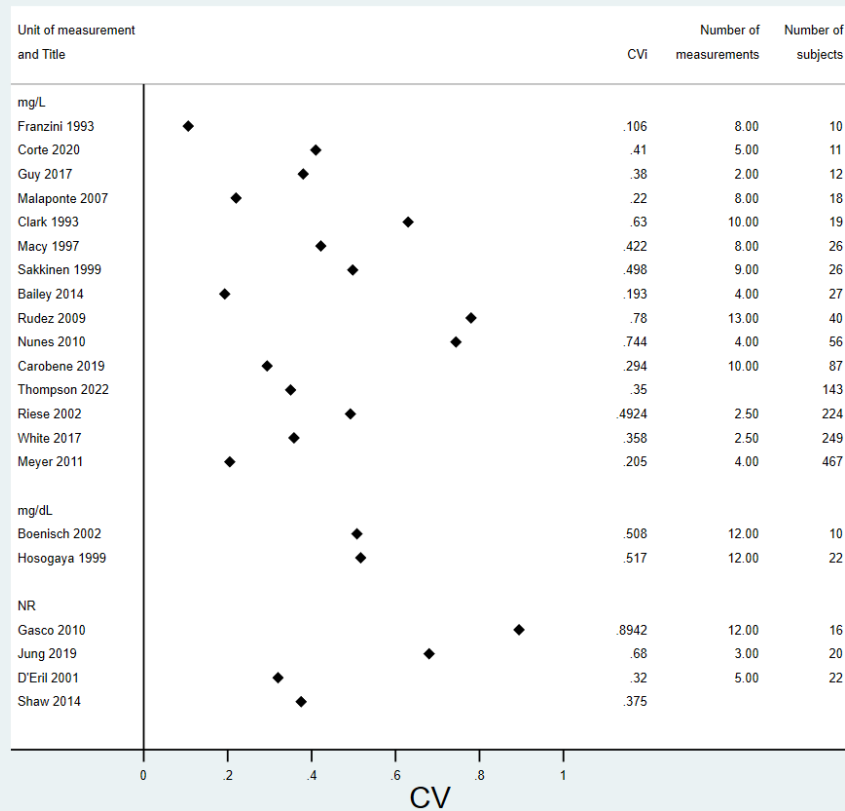


Figure 13.4 Forest plot of all studies that reported a CV for CRP, split by unit of measurement

13.4.2 hsCRP CVs

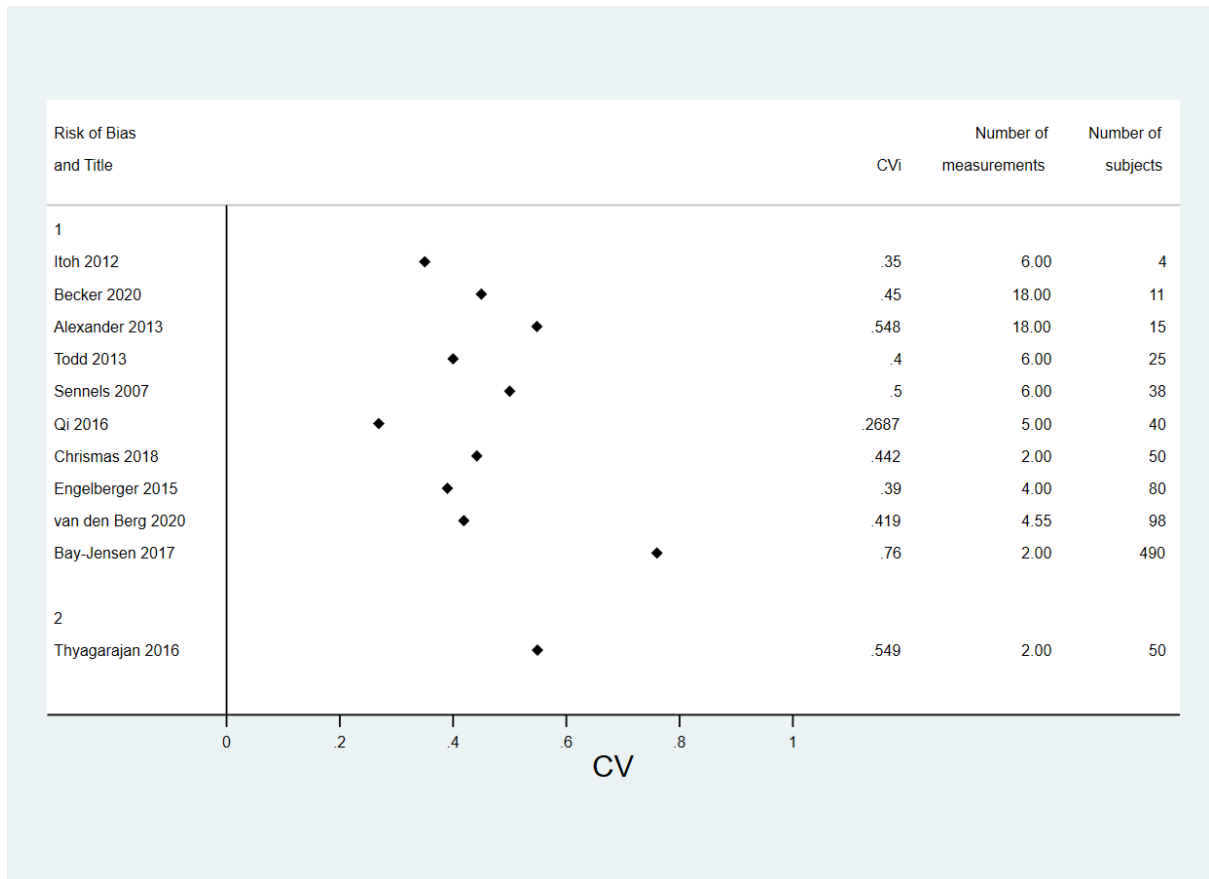


Figure 13.5 Forest plot of all studies that reported a CV for hsCRP, split by Risk of Bias

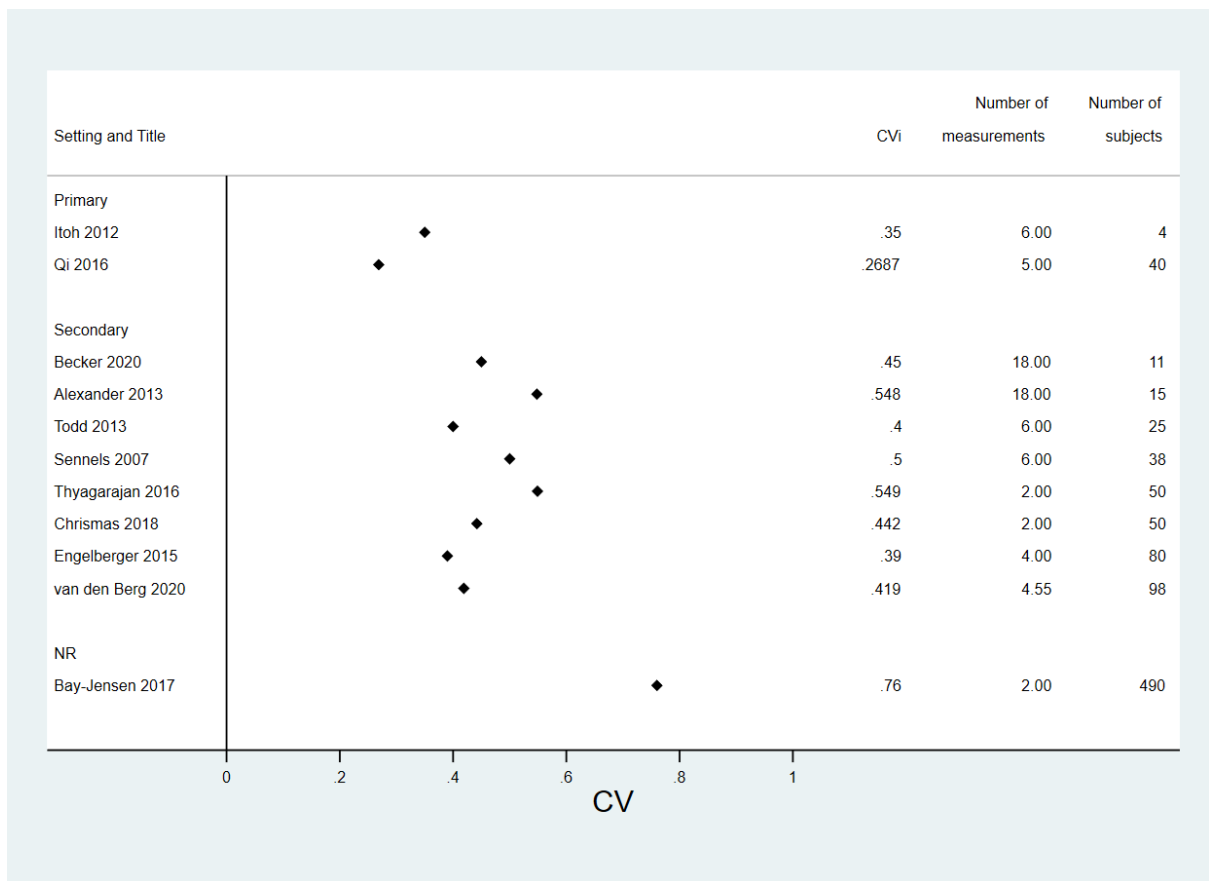


Figure 13.6 Forest plot of all studies that reported a CV for hsCRP, split by setting

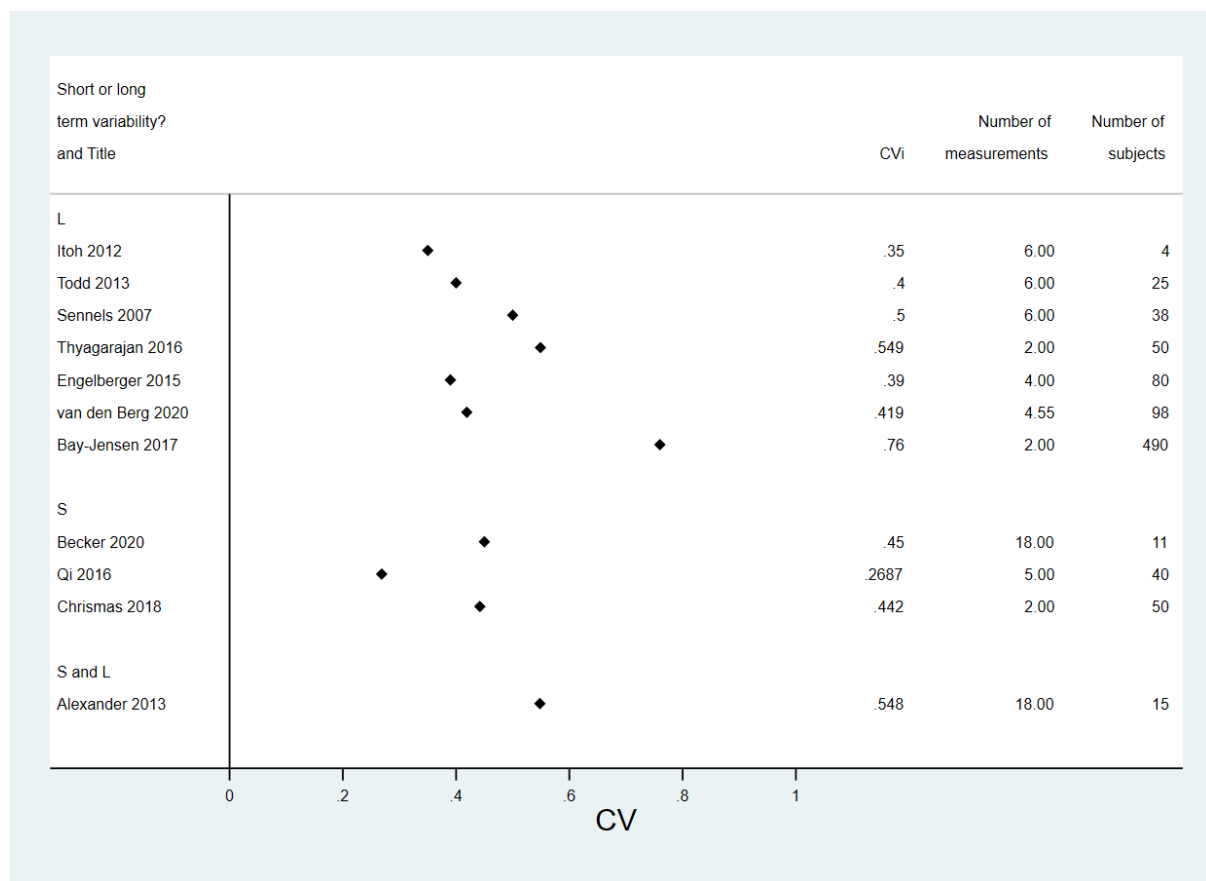


Figure 13.7 Forest plot of all studies that reported a CV for hsCRP, split by short (S) or long (L) term measurement of variation

13.4.3 CRP ICCs

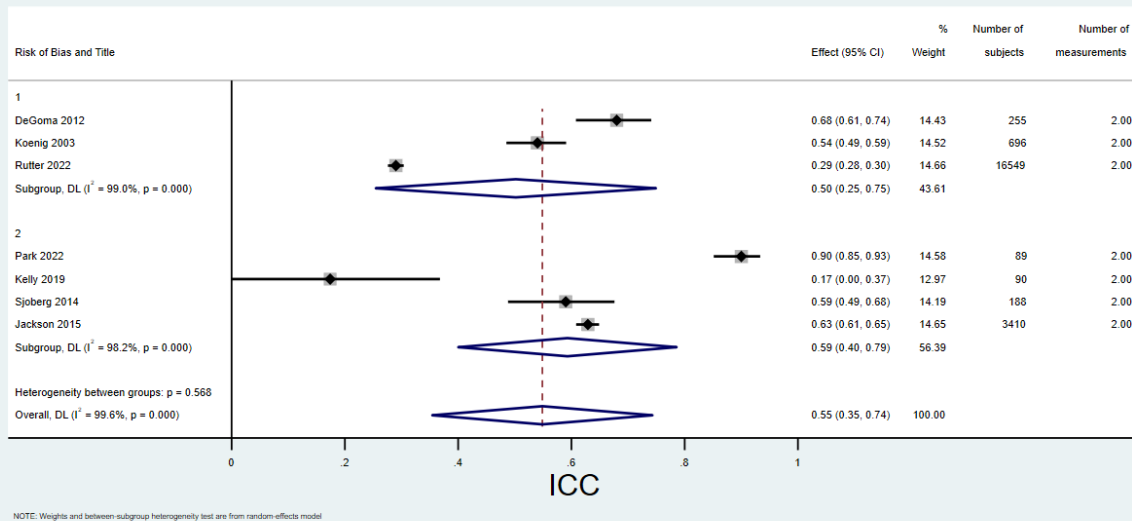


Figure 13.8 Forest plot of all studies that reported an ICC for CRP, split by Risk of Bias, with meta-analysis

13.4.4 hsCRP ICCs

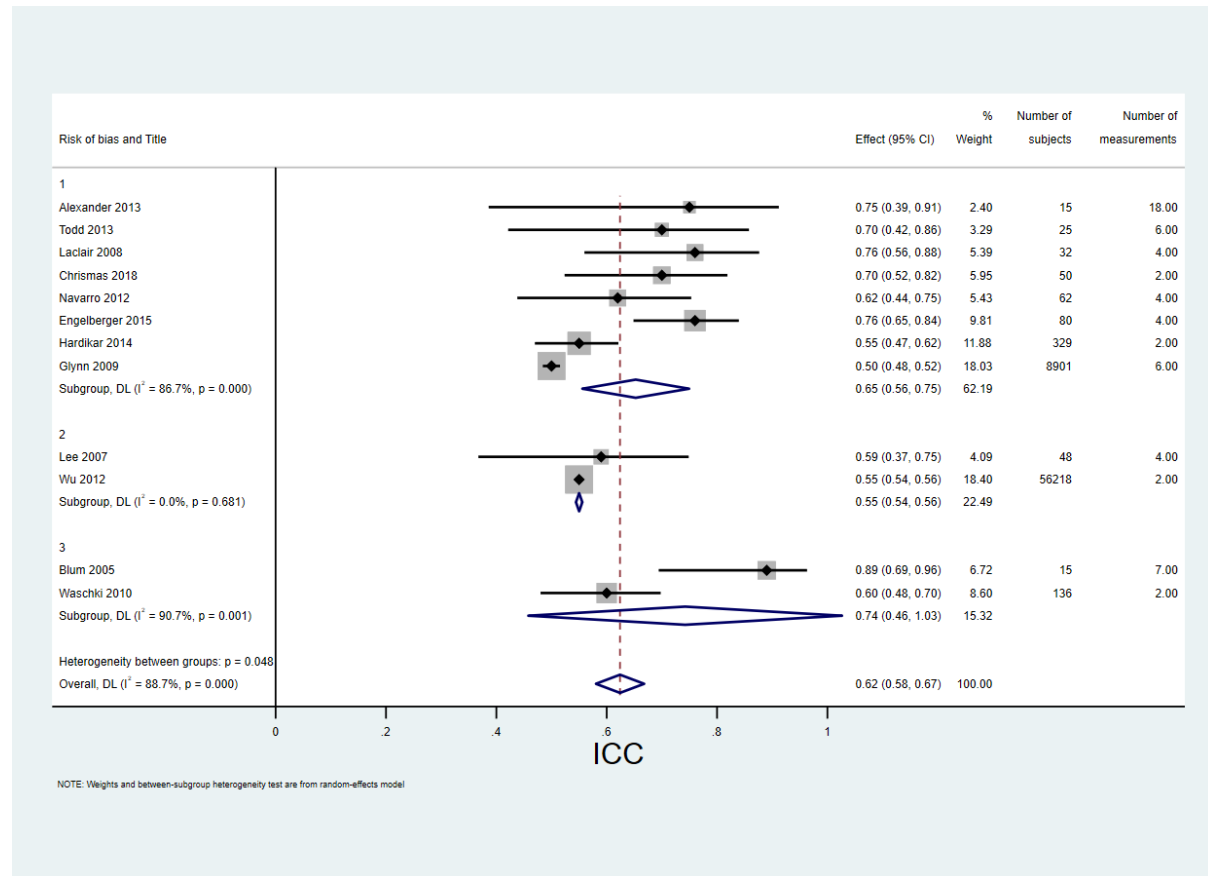


Figure 13.9 Forest plot of all studies that reported an ICC for hsCRP, split by Risk of Bias, with meta-analysis

14 Appendix 4 Spirometry systematic review

appendices

14.1 Appendix 4.1 Search strategies

14.1.1.1 A. (Medline, Embase, Medline and In-Process, In-Data-Review & Other non-indexed citations)

1. spirometry.mp. or Spirometry/ or lung function.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
2. Fev1.mp. or Forced Expiratory Volume/
3. FVC.mp. or Respiratory Function Tests/ or Vital Capacity/
4. biological varia\$.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
5. within-individual varia\$.mp.
6. within-subject varia\$.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
7. within-individual CV.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
8. Within-subject CV.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
9. Intraindividual CV.mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]

10. intra-individual varia\$.mp.
11. day-to-day.mp.
12. visit-to-visit.mp.
13. Coefficient of variation.mp.
14. (cvi or cvw or cvb or cvwithin-subject or cvbiological).mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
15. chronic venous insufficiency.mp.
16. vaccine immunology.mp.
17. content validity index.mp.
18. cerebral visual impairment.mp.
19. community virtual ward.mp.
20. cardiac vagal withdrawal.mp.
21. coronary vessel wall.mp.
22. variability independent of the mean.mp.
23. VIM.mp.
24. thalamus.mp. or Thalamus/
25. lactamase.mp. or beta-Lactamases/ or beta-Lactamase Inhibitors/
26. index of individuality.mp.
27. Reference Change Value.mp.
28. rcv.mp.

29. (intraclass or ICC).mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
30. Revue canadienne du vieillissement.mp.
31. recoverin.mp. or Recoverin/
32. index of heterogeneity.mp.
33. validity coefficient.mp.
34. reliability parameter.mp.
35. 1 or 2 or 3
36. (animal not human).mp. [mp=ti, ab, hw, tn, ot, dm, mf, dv, kw, fx, dq, nm, kf, ox, px, rx, ui, sy]
37. 15 or 16 or 17 or 18 or 19 or 20 or 21 or 24 or 25 or 30 or 31
38. 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 22 or 23 or 26 or 27 or 28 or 29 or 32 or 33 or 34
39. 38 not 37
40. 35 and 39
41. 40
42. remove duplicates from 41

14.1.1.2 B. Cochrane central

FEV1 or FVC or spirometry

AND

variability or variation or within-individual or within-subject or day-to-day or test-retest

14.1.1.3 C. *Epistemonikos*

(title:((FEV1 OR FVC OR SPIROMETRY) AND variability) OR abstract:((FEV1 OR FVC OR SPIROMETRY) AND variability))

D. Google Scholar

(FEV1 or FVC or spirometry) and variability

14.2 Appendix 4.2 Variables extracted from studies

Table 14.1 Variables extracted from studies

Variable	Definition and rules
Spirometry measurand	FEV1 or FVC
Study design	Cohort, RCT etc
Number of individuals	Number of individuals used to calculate variation measure
Age	Average age of individuals used to calculate variation measure. If the individuals used to calculate variation measure is a subset of the study population and only the age is given for the whole population, the age of the whole population is used.
Sex	Percentage of individuals used to calculate variation measure who were male. If the individuals used to calculate variation measure is a subset of the study population and only the sex is given for the whole population, the sex of the whole population is used.
Ethnicity	Ethnicity recorded in study.
Setting	Primary care/community or secondary/tertiary/laboratory setting
Health status	1=healthy; 2=COPD; 3=asthma; 4=cystic fibrosis 5=other
Number of measurements	Number of repeated applications of depression measurement instrument to the same individual
Time interval between measurements	Length of time between measurements. Categorised by within-day, short-term (<7 days) and long-term (>7 days)
Unit of measurements	L, mL or %pred
CV _i	Coefficient of variation of repeated measures within an individual
SD	Standard deviation of repeated measures within an individual
ICC	Intraclass correlation of repeated measures within an individual

14.3 Appendix 4.3 Forest plots of subgroup analyses

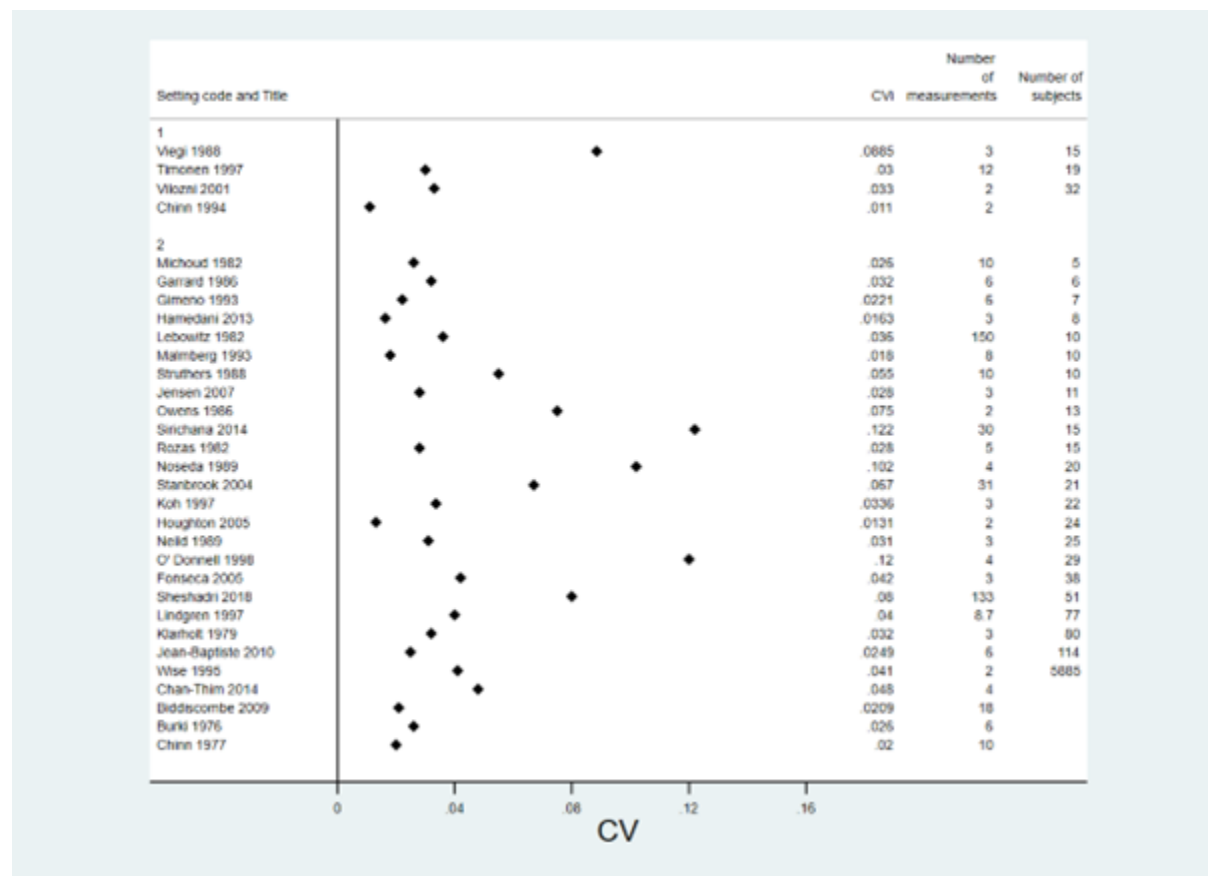


Figure 14.1 Forest plot of within-individual CV of FEV1 by setting, all studies using units of Litres.

Setting code 1=primary or community care, 2=secondary or tertiary care.

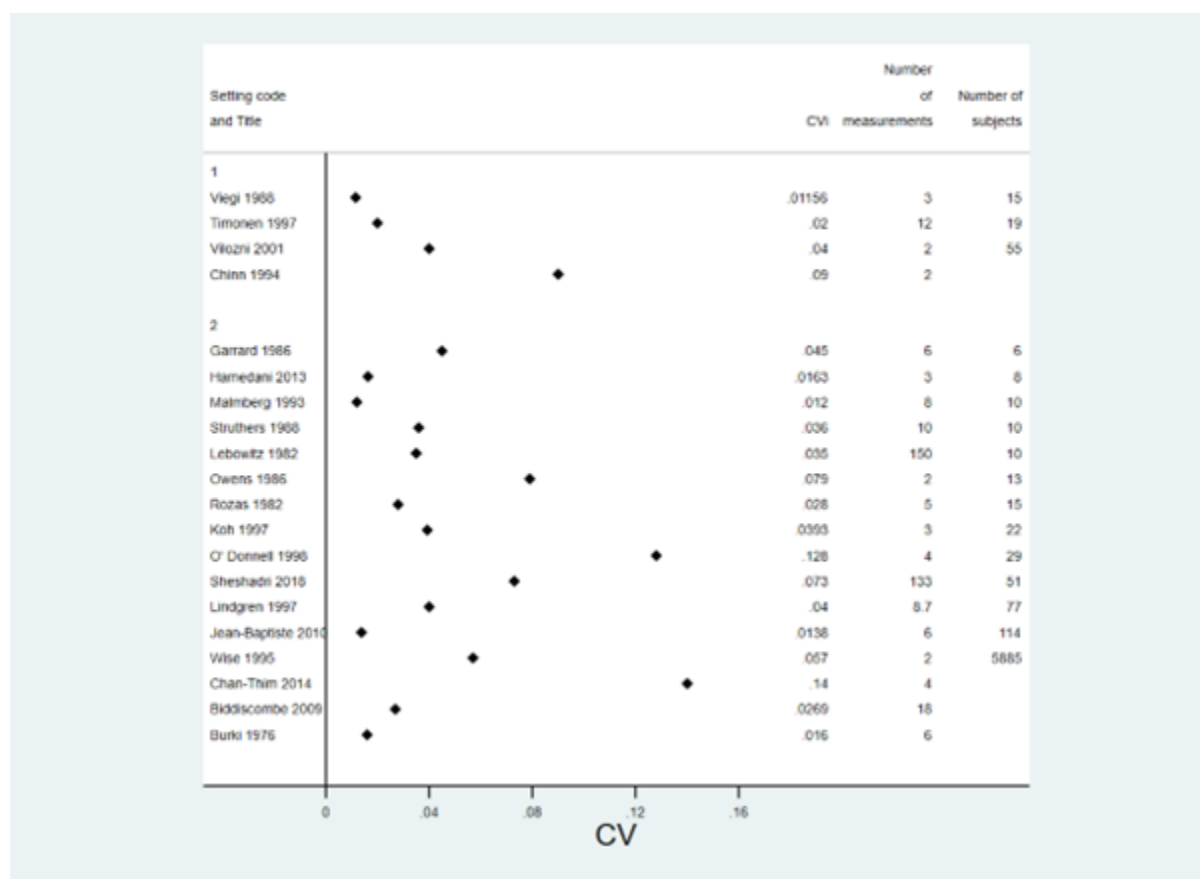


Figure 14.2 Forest plot of within-individual CV of FVC by setting, all studies using units of Litres.

Setting code 1=primary or community care, 2=secondary or tertiary care.

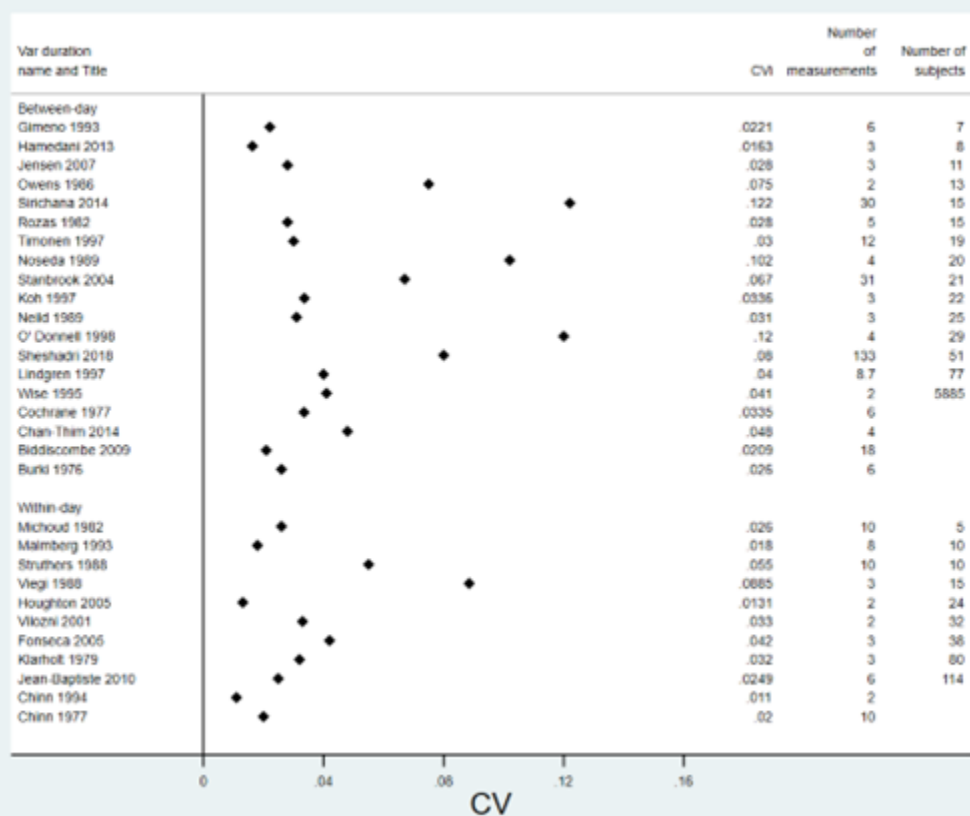


Figure 14.3 Forest plot of within-individual CV of FEV1 by within or between-day, all studies using units of Litres.

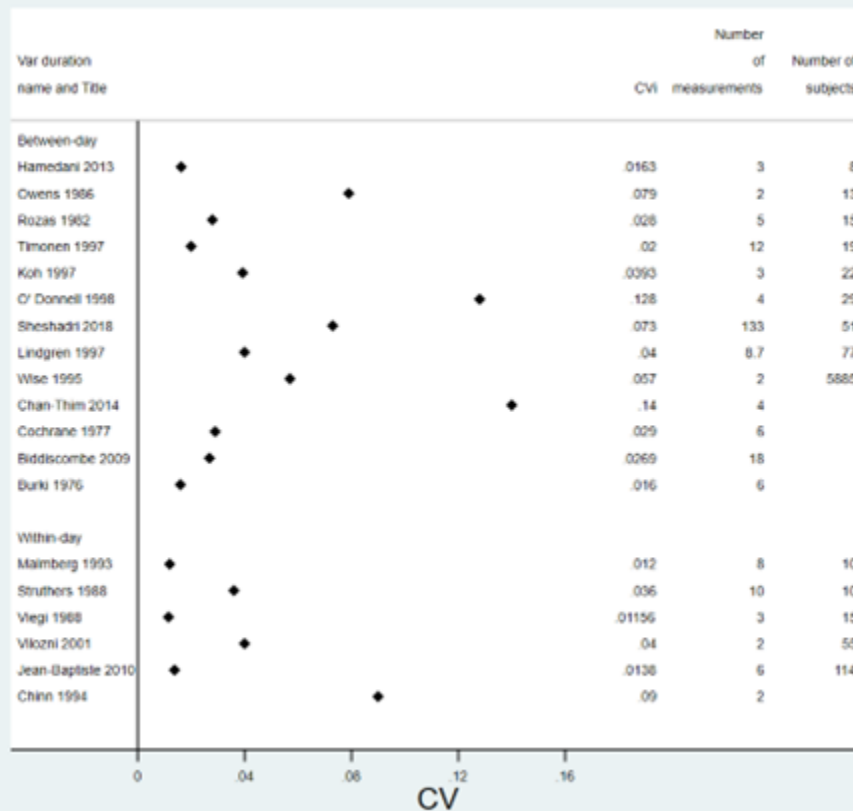


Figure 14.4 Forest plot of within-individual CV of FVC by within or between-day, all studies using units of Litres.

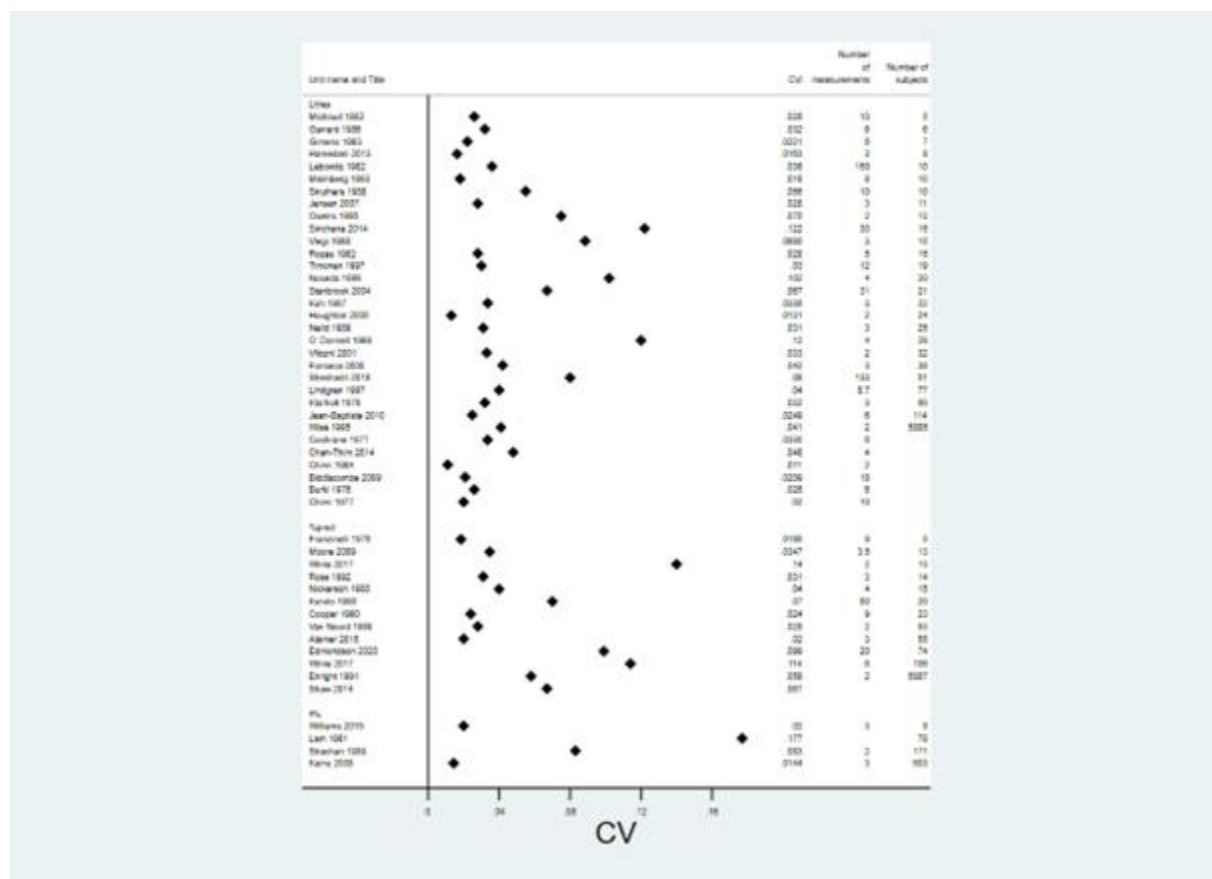


Figure 14.5 Forest plot of within-individual CV of FEV1 by unit

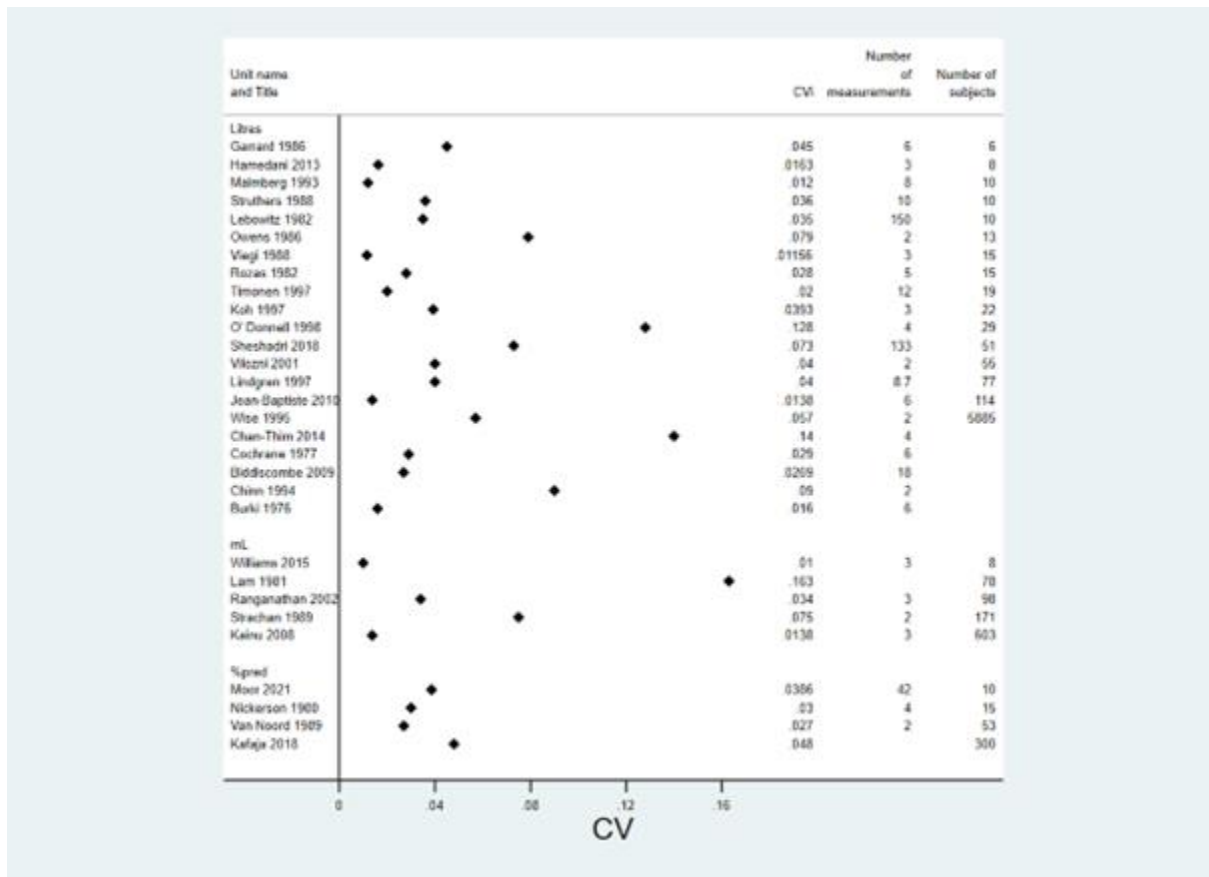


Figure 14.6 Forest plot of within-individual CV of FVC by unit

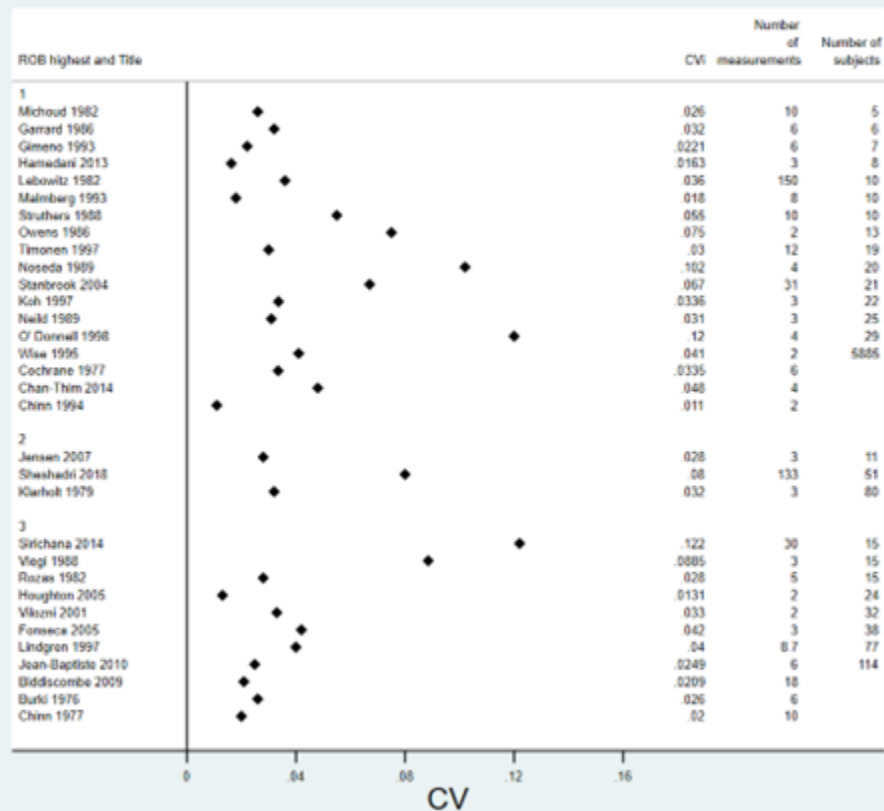


Figure 14.7 Forest plot of within-individual CV of FEV1 by highest risk of bias score, all studies using units of Litres

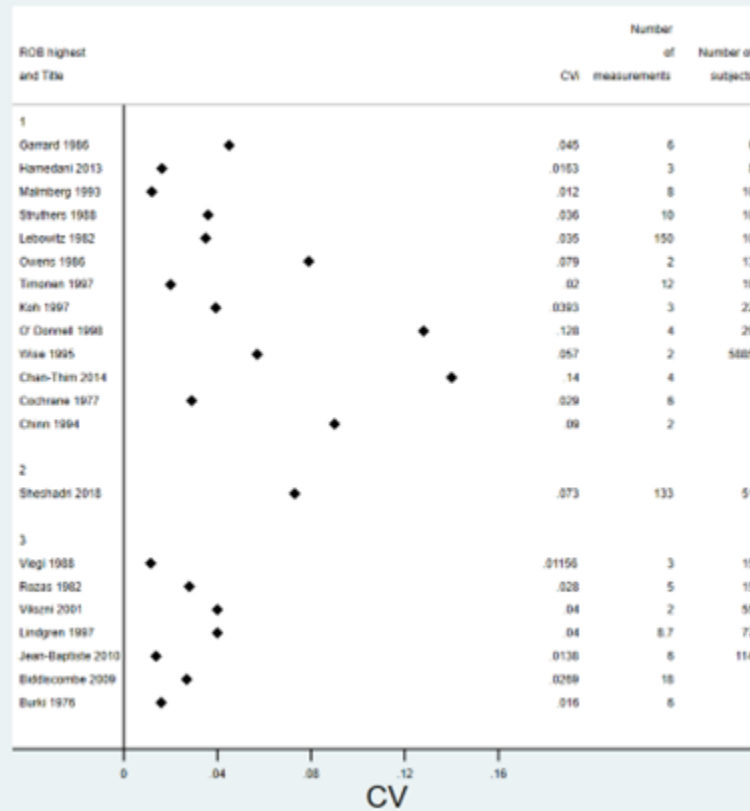


Figure 14.8 Forest plot of within-individual CV of FVC by highest risk of bias score, all studies using units of Litres

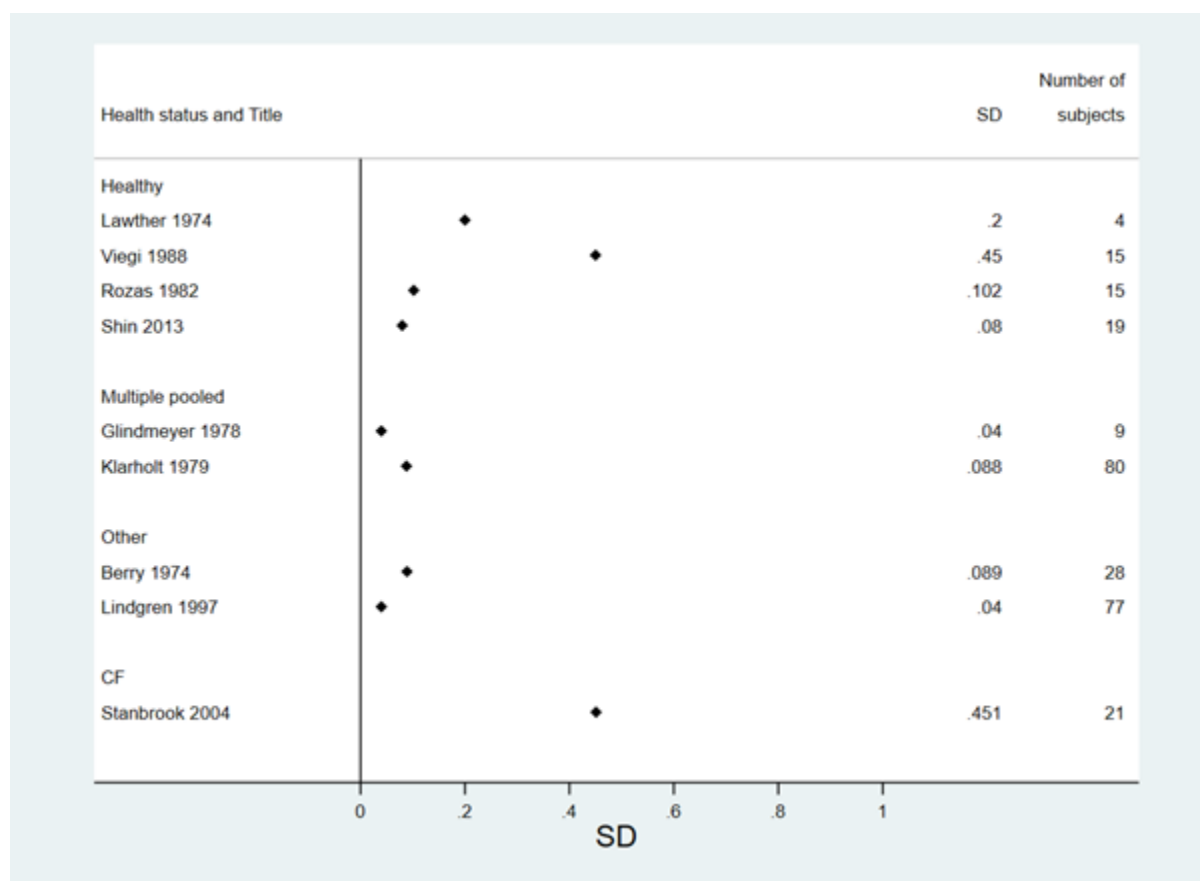


Figure 14.9 Forest plot of within-individual SD of FEV1 by health status, all studies using units of Litres

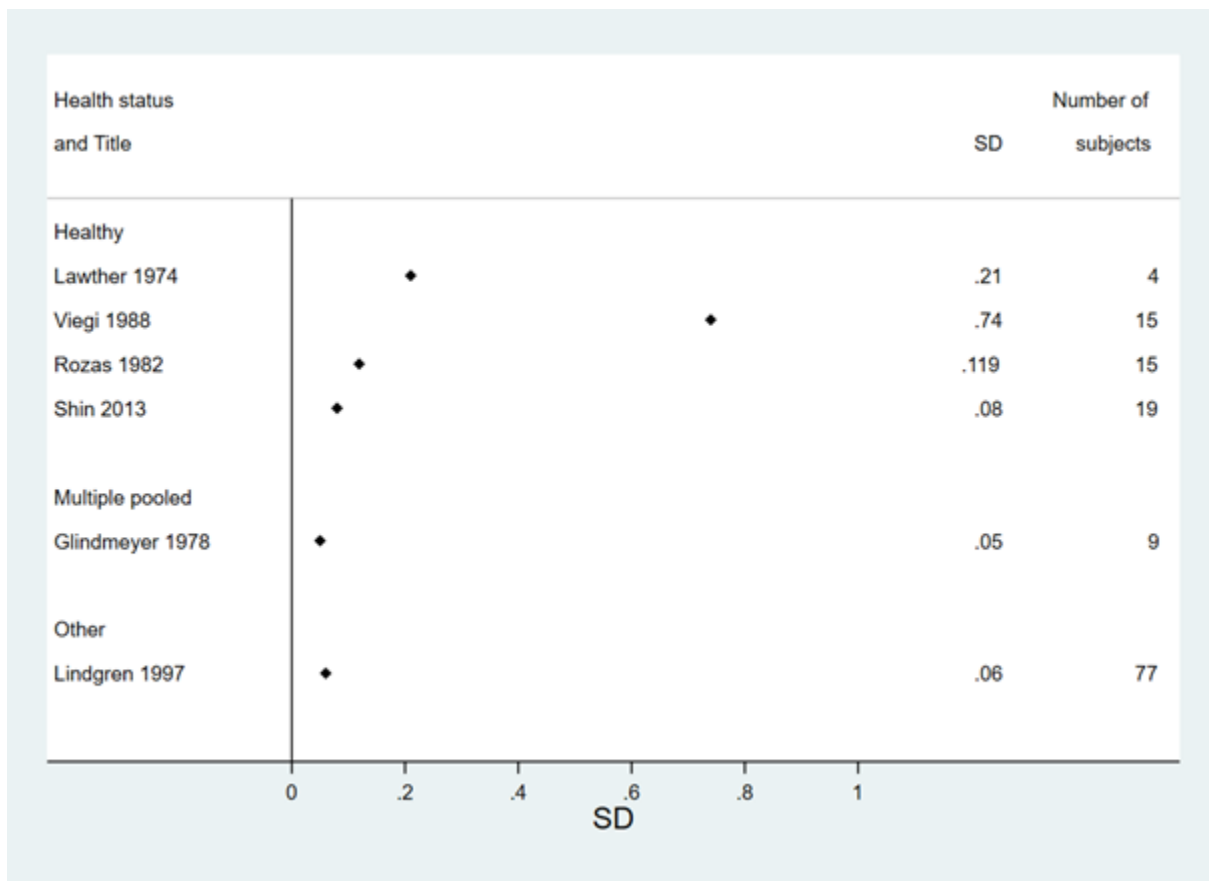


Figure 14.10 Forest plot of within-individual SD of FVC by health status, all studies using units of Litres

15 Appendix 5 Depression symptom measurement

instrument systematic reviews

15.1 Description of Depression symptom measurement instruments

15.1.1.1.1 1. Patient health questionnaire-9 (PHQ-9)

The PHQ-9 is a nine question, self-administered questionnaire, and assesses the nine DSM-IV criteria for depression using the depression module from the fuller Patient Health Questionnaire. If five or more of the criteria are positive for at least half of the days of the past two weeks, then major depression is suspected. Each question can be scored one to three, giving a total possible score of 27. Higher scores are associated with higher severity of depression. It has been validated in a primary care setting against the Mental Health Personnel re-interview, with a score of ten or more giving sensitivity of 88% and specificity of 88%. The minimal clinically important difference (MCID) has been estimated as 5 points [368, 578] In a research study in primary care, a remission was defined as score of less than 5 after 6 months of treatment, and a response was defined as a 50% reduction in score [579] Another study defined an adequate response as a 5 point fall in PHQ-9 score after six months of treatment, with a 2-4 point fall being borderline [580].

15.1.1.1.2 2. Patient health questionnaire-2 (PHQ-2)

The PHQ-2 is a shortened, very brief version of the PHQ-9. Originally it was recommended that a score of 3 or more out of a possible 6 was used as the cut-off point to diagnose major depression, but a more recent meta-analysis has

recommended a cut-off point of 2 to improve sensitivity. It has been validated in a primary care and obstetrics-gynaecology setting against the Mental Health Personnel re-interview to give a sensitivity of 83% and a specificity of 92% for a cut-off point of 3 or more. The meta-analysis, which considered studies from a variety of settings, found a sensitivity of 91% and a specificity of 70% when a cut-off point of 2 or more was used, as compared to a “recognised gold-standard instrument” using either DSM or International Classification of Disease (ICD) criteria for major depression. A value for MCID was not found in the literature, possibly because there are only a small number of possible scores, and because the PHQ-2 is intended for screening for diagnosis, not for monitoring (the PHQ-9 is preferred to assess response to treatment) [581, 582].

15.1.1.1.3 3. Hospital Anxiety and Depressions Scale (HADS)

The HADS instrument comprises 14 questions, 7 regarding anxiety and 7 regarding depression. The depression questions focus on anhedonia. Each question can be scored 0 to 3, for a maximum score of 21 for each of anxiety and depression. A score greater than or equal to 11 is suggestive of a diagnosis of anxiety or depression with a score of 8 to 10 equivocal. It has been validated in general practice using a structured clinical interview based on DSM-3 criteria, giving a sensitivity of 90% and a specificity of 86%. Note that in the same study, general practitioners were asked to assess, blind to the HADS or structured clinical interview, whether the patient had physical, psychological or a mixture of symptoms, and achieved a sensitivity of 49% but a specificity of 96%. The MCID for the HADS depression subscale was estimated to be 2 points in a group of patients with bronchiectasis [583-585].

15.1.1.1.4 4. Beck Depression Inventory first and second editions (BDI and BDI-II)

The Beck Depression Inventory aimed to assess the presence and severity of depression in both adults and adolescents. It was updated to the second edition to make it more compatible with DSM criteria. Both versions of the test have 21 self-reported questions which can each be scored from 0 to 3, making a total possible score of 63. A score of 14-19 suggests mild depression, 20-28 moderate depression, and 29 or above severe depression. A meta-analysis of studies in a variety of settings showed an optimal cut-off point of 14.5 leading to a sensitivity of 86% and a specificity of 78%. In most of the included studies the reference standard was a structured clinical interview. The MCID has been estimated as a reduction in score of 17.5% from baseline [371, 384, 586, 587].

15.2 Appendix 5.2 Databases searched

Table 15.1 Databases searched

Database	Coverage
Ovid	
Medline	1946 to 14/7/23
Embase	1974 to 14/7/23
APA Psychinfo	1967 to 14/7/23
Medline and In-Process, In-Data-Review & Other non-indexed citations	1946 to 14/7/23
Cochrane Central	Inception to 14/7/23
Epistemonikos	Inception to 14/7/23
Open Grey	Inception to 14/7/23

15.3 Appendix 5.3 Search strategies

15.3.1.1 A. (*Medline, Embase, APA Psycinfo, Medline and In-Process, In-Data-Review & Other non-indexed citations*)

1. (Beck depression inventory or BDI).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
2. ((Hospital Anxiety and Depression Score) or HADS).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
3. (Patient health Questionnaire-9 or PHQ-9 or Patient health Questionnaire-2 or PHQ-2).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
4. biological varia\$.mp.
5. (within-individual varia\$ or within-subject varia\$ or intra-individual varia\$).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

6. (within-individual CV or within-subject CV or Intraindividual CV).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
7. (day-to-day or visit-to-visit).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
8. (Coefficient of variation or cvi or cvw or cvb or cvwithin-subject or cvbiological).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
9. (variability independent of the mean or VIM).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
10. (index of individuality or Reference Change Value or RCV or intraclass or ICC).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

11. (index of heterogeneity or validity coefficient or reliability parameter or Cohen's kappa or Cronbach's alpha or agreement).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

12. 1 or 2 or 3

13. 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11

14. 12 and 13

15. remove duplicates from 14

15.3.1.2 B. Cochrane central

(Patient health questionnaire) or (Becks Depression Inventory) or (Hospital anxiety and Depression scale)

AND

variability or variation or within-individual or within-subject or day-to-day or test-retest

15.3.1.3 C. Epistemonikos

(title:((PHQ-9 OR PHQ-2 OR HADS OR BDI) AND variability) OR abstract:((PHQ-9 OR PHQ-2 OR HADS OR BDI) AND variability))

15.3.1.4 D. Open Grey

(PHQ-9 OR HADS OR BDI) and variability

15.4 Appendix 5.4 Included studies and characteristics

Paper	PHQ-9, PHQ-2, BDI, HADS (or multiple)	Number of subjects	Age (primary analysis)	Sex %male (primary analysis)
Bakalidou 2013 [588]	BDI-II	147		
Bakalidou 2013 [588]	BDI-II	72	43.17	33.33
Bakalidou 2013 [588]	BDI-II	75	38.83	32.00
Beswick 2022 [311]	PHQ-9	102	50	23
Boyer 2011 [589]	HADS - depression subscale only	35		31.43
Chen 2016 [590]	PHQ-9	82		
Chiu 2022 [591]	BDI-II	59	55.3	66.1
Delgadillo 2011 [312]	PHQ-9	60	35	77.00
Delgadillo 2011 [312]	PHQ-2	60	35	77.00
Duan-Porter 2018 [310]	PHQ-9	281	41.4	7.80
Ganguly 2013 [592]	PHQ-9	233	15.89	53.64
Gelaye 2014 [593]	PHQ-9	46	35	39.00
Hansen 2020 [594]	HADS-D	50	67	56
Hitchon 2019 [595]	PHQ-9	115	59.8	15.30
Hitchon 2019 [595]	HADS-D	115	59.8	15.30
Huang 2017 [596]	BDI-II	70	67.7	62.86
Hyland 2019 [313]	HADS	242	51.7	0.00
Jin 2020 [597]	PHQ-2 (SMS)	53	55.35	35.80
Jin 2020 [597]	PHQ-2 (INT)	52	57.33	50.00
Karmisholt 2019 [306]	HADS	15	57.5	10.00
Kha 2020 [598]	PHQ-9	99	47.3	25.25
Lamers 2008 [599]	PHQ-9	106	71.4	51.80
Lamers 2008 [599]	PHQ-9	27		
Lamers 2008 [599]	PHQ-9	79		
Li 2015 [600]	PHQ-9	50	60.8	53.70
Li 2015 [600]	PHQ-2	50	60.8	53.70
Lodhi 2020 [601]	HADS	200		0.00
Lowe 2004 [309]	PHQ-9	82	70	27.00
Lowe 2004 [309]	PHQ-9	41	71.8	32.00
Marinus 2002 [602]	HADS	54	65.2	56.00
Michopoulos 2008 [603]	HADS	50	37.11	59.30
Mohamadian 2022 [604]	PHQ-9	246	20.43	49.6
Mohamadian 2022 [604]	PHQ-2	246	20.43	49.6
Monohan 2009 [605]	PHQ-9	187	36	27.00
Omoro 2006 [606]	PHQ-9	24	55.7	91.67
Pinto-Meza 2005 [607]	PHQ-9 by telephone	58	50.84	15.50
Pinto-Meza 2005 [607]	PHQ-9 self-administered	57	38.7	33.30
Quek 2001 [314]	BDI	108		
Rae 2020 [608]	HADS	86		55.65
Reda 2011 [609]	HADS	144	33.8	34.40
Shelton 2015 [307]	HADS-D	23	25.4	30.00
Silverberg 2021 [610]	PHQ-9	458	44.4	40.8
Silverberg 2021 [610]	PHQ-2	458	44.4	40.8
Summaka 2019 [611]	PHQ-9	20	37.2	100.00
Tanveer Hasan 2017 [612]	PHQ-9	130		
Tiksnadi 2023 [317]	HADS	41	42.41	45.5
Visser 2006 [613]	BDI	50	65	62.00
Wang 2009 [614]	HADS	173	59	79.90
Weobong 2009 [308]	PHQ-9	40	27.1	0.00

Paper	PHQ-9, PHQ-2, BDI, HADS (or multiple)	Number of subjects	Age (primary analysis)	Sex %male (primary analysis)
Woldetensay 2018 [615]	PHQ-9	29	24.3	0.00
Yu 2011 [616]	PHQ-2	203	15	46.20
Yu 2012 [617]	PHQ-9	203		46.20
Yusoff 2011 [315]	HADS-D subscale	67	51	100.00

Characteristics of studies included

Paper	Setting code	Health status code	Number of measurements	Time between measurements (days)	Variation measure	Risk of Bias score
Bakalidou 2013	2	5	2	7	ICC	3
Bakalidou 2013	2	4	2	7	ICC	3
Bakalidou 2013	2	1	2	7	ICC	3
Beswick 2022	2	4	3	14	ICC	1
Boyer 2011		4	2	18	ICC	2
Chen 2016	1	1	2	14	ICC	2
Chiu 2022	2	4	2	14	ICC	1
Delgadillo 2011	1	3	2	28	ICC	2
Delgadillo 2011	1	3	2	28	ICC	2
Duan-Porter 2018	2	1	12	30	ICC	2
Ganguly 2013	1	1	2	30	ICC	2
Gelaye 2014	2	4	2	1	ICC	3
Hansen 2020	2	4	20	10	ICC	1
Hitchon 2019	2	4	2	14	ICC	3
Hitchon 2019	2	4	2	14	ICC	3
Huang 2017	2	4	2	14	ICC	2
Hyland 2019	2	4	2	42	ICC	2
Jin 2020	1	1	2	7	ICC	3
Jin 2020	1	1	2	7	ICC	3
Karmisholt 2019	2	4	13	30	CV _i	1
Kha 2020	2	4	2	14	ICC	1
Lamers 2008	1	4	2	7.9	ICC	2
Lamers 2008	1	4	2	7.9	ICC	2
Lamers 2008	1	4	2	7.9	ICC	2
Li 2015	2	4	2	7	ICC	2
Li 2015	2	4	2	7	ICC	2
Lodhi 2020	2	4	2	14	ICC	2
Lowe 2004	2	5	2		ICC	3
Lowe 2004	2	5	2		ICC	3
Marinus 2002	2	4	2	14	ICC	2
Michopoulos 2008	2	1	2	20	ICC	2
Mohamadian 2022	2	1	2	14	ICC	1
Mohamadian 2022	2	1	2	14	ICC	1
Monohan 2009	2	4	2	7	ICC	2
Omoro 2006	2	4	2	14	ICC	2
Pinto-Meza 2005	1	4	2	7	ICC	3
Pinto-Meza 2005	1	4	2	7	ICC	3
Quek 2001	2	4	2	84	ICC	2
Rae 2020	2	4	2		ICC	3
Reda 2011		4	2	14	ICC	3
Shelton 2015	2	4	2	42	ICC	3
Silverberg 2021	2	4	2	146	ICC	1
Silverberg 2021	2	4	2	146	ICC	1
Summaka 2019	2	4	2	15	ICC	3
Tanveer Hasan 2017	2	4	2	7	ICC	3
Tiksnadi 2023	1	1	2	90	ICC	1
Visser 2006	2	4	2	14	ICC	3
Wang 2009	2	4	2	14	ICC	3
Weobong 2009	1	1	2	14	ICC	3
Woldetensay 2018	1	4	2	7	ICC	3

Yu 2011	1	1	2	7	ICC	3
Yu 2012	1	1	2		ICC	3
Yusoff 2011	2	1	2	42	ICC	3

Characteristics of studies included (cont.)

Setting code: 1=primary or community;2=tertiary/secondary/laboratory

Health status code: 1=Healthy; 2=depression 3=other mental health condition 4=physical health condition 5=mixed

Risk of bias are presented as highest score, with questions 4 and 5 excluded. 1=very good; 2=adequate; 3=doubtful; 4=inadequate

15.5 Appendix 5.5 Risk of bias scores for all studies

Paper	Q1	Q2	Q3	Q6	Q7	Overall score
Bakalidou 2013	1	1	1	1	3	3
Beswick 2022	1	1	1	1	1	1
Boyer 2011	1	1	2	2	2	2
Chen 2016	2	1	2	1	2	2
Chiu 2022	1	1	1	1	1	1
Delgadillo 2011	1	1	1	1	2	2
Duan-Porter 2018	1	1	1	2	1	2
Ganguly 2013	1	1	1	1	2	2
Gelaye 2014	2	3	1	1	2	3
Hansen 2020	1	1	1	1	1	1
Hitchon 2019	2	3	1	1	2	3
Huang 2017	1	1	1	1	2	2
Hyland 2019	2	2	1	1	2	2
Jin 2020	1	1	1	1	3	3
Karmisholt 2019	1	1	1	1	1	1
Kha 2020	1	1	1	1	1	1
Lamers 2008	1	1	1	1	2	2
Li 2015	2	2	2	2	2	2
Lodhi 2020	2	1	1	1	2	2
Lowe 2004	2	1	1	1	3	3
Marinus 2002	2	1	1	1	2	2
Michopoulos 2008	1	1	1	1	2	2
Mohamadian 2022	1	1	1	1	1	1
Monohan 2009	2	1	1	1	1	2
Omoro 2006	2	1	2	1	2	2
Pinto-Meza 2005	2	3	2	1	1	3
Quek 2001	1	2	2	1	1	2
Rae 2020	2	3	3	1	3	3
Reda 2011	2	1	1	1	3	3
Shelton 2015	3	2	2	1	2	3
Silverberg 2021	1	1	1	1	1	1
Summaka 2019	1	1	1	1	3	3
Tanveer Hasan 2017	3	2	3	1	2	3
Tiksnadi 2023	1	1	1	1	1	1
Visser 2006	1	1	1	1	3	3
Wang 2009	2	1	1	1	3	3
Weobong 2009	2	1	2	1	3	3
Woldetensay 2018	2	1	2	1	3	3
Yu 2011	2	1	1	1	3	3
Yu 2012	2	1	1	1	3	3
Yusoff 2011	3	2	1	1	2	3

15.6 Appendix 5.6 Heterogeneity results

Subgroup	Tau-squared	Predictive interval
PHQ-9	0.005	0.69-1.00
HADS	0.003	0.76-1.02
BDI	0.002	0.73-1.03
PHQ-2	0.001	0.56-0.93

Heterogeneity by PROM

Subgroup	Tau-squared	Predictive interval
Primary	0.006	0.66-1.06
Secondary/other	0.004	0.69-0.98

Heterogeneity by setting

Subgroup	Tau-squared	Predictive interval
Healthy	0.004	0.59-0.98
Physical health condition	0.004	0.74-1.02
Mental health condition	0	0.68-0.88
Mix	0	0.73-0.89

Heterogeneity by health condition

Tau-squared	Predictive interval
0.004	0.68-0.98

Heterogeneity where number of subjects >100

Heterogeneity (I^2) %	Predictive interval
0.903	0.72-0.94

Heterogeneity where time between measurements >7 days

Tau-squared	Predictive interval
0.004	0.68-0.96

Heterogeneity where risk of bias <3

16 Appendix 6 HbA1c cohort study appendices

16.1 Appendix 6.1 - Subgroup analysis

Table 16.1 Other subgroup analyses

		N	%	Mean (mmol/mol)	CV (95%CI)	95% LCI	95% UCI
Comorbidities	No comorbidity (incl DM)	36,157	6.2	37.52	0.102	0.101	0.102
	PDM, DM1 or DM2 with no comorbidity	66,710	11.4	61.71	0.189	0.188	0.189
	Hypertension	324,047	55.2	52.60	0.194	0.194	0.195
	Asthma	102,976	17.5	52.38	0.203	0.203	0.204
	COPD	47,079	8.0	52.31	0.200	0.199	0.200
	Hyperthyroidism	12,167	2.1	51.30	0.193	0.192	0.194
	Hypothyroidism	65,616	11.2	51.81	0.195	0.194	0.196
	CKDstage3 to 5	97,611	16.6	53.64	0.197	0.196	0.197
	Heart failure	22,740	3.9	55.51	0.206	0.205	0.207
	Heart disease	102,420	17.4	53.01	0.190	0.189	0.190
	Ischaemic stroke	12,453	2.1	53.70	0.199	0.198	0.200
	Haem. Stroke	3,926	0.7	52.06	0.204	0.201	0.206
	Cancer	50,146	8.5	51.91	0.190	0.189	0.190
	PDM, DM1 or DM2 with 1 or more comorbidity	255,798	43.6	57.68	0.185	0.184	0.185
	1 or more comorbidity with no PDM, DM1 or DM2	80,444	13.7	38.55	0.087	0.086	0.087
Diabetic medication	None	253,534	43.2	40.47	0.077	0.077	0.077
	On diabetic drug but not insulin	207,239	35.3	57.39	0.184	0.184	0.184
	Insulin	81,420	13.9	69.66	0.170	0.170	0.170
	All diabetic drugs	299,000	50.9	61.39	0.190	0.190	0.190
Region	East Midlands	3,815	0.6	59.04	0.164	0.162	0.166
	East of England	17,745	3.0	54.48	0.186	0.185	0.187
	London	62,903	10.7	51.29	0.196	0.196	0.197
	North East	7,255	1.2	55.31	0.184	0.182	0.185
	North West	87,175	14.9	47.55	0.188	0.188	0.189
	Northern Ireland	38,782	6.6	50.36	0.209	0.208	0.210
	Scotland	87,135	14.8	57.68	0.206	0.206	0.207
	South Central	31,671	5.4	57.84	0.192	0.192	0.193
	South East Coast	49,530	8.4	53.38	0.198	0.198	0.199
	South West	31,404	5.3	54.42	0.191	0.190	0.192
	Wales	111,820	19.0	50.25	0.204	0.203	0.204
	West Midlands	53,147	9.1	50.67	0.183	0.182	0.183
	Yorkshire & Humber	4,641	0.8	55.72	0.190	0.188	0.192
Alcohol consumption	Current drinker	344,894	58.8	51.73	0.197	0.197	0.198
	Ex drinker	10,562	1.8	54.68	0.218	0.217	0.220
	Teetotal	119,829	20.4	53.53	0.202	0.201	0.202
Smoker Status	Current Smoker	117,878	20.1	53.58	0.214	0.213	0.214
	Ex-smoker	127,989	21.8	52.63	0.196	0.196	0.197
	Never smoked	278,076	47.4	51.72	0.195	0.194	0.195

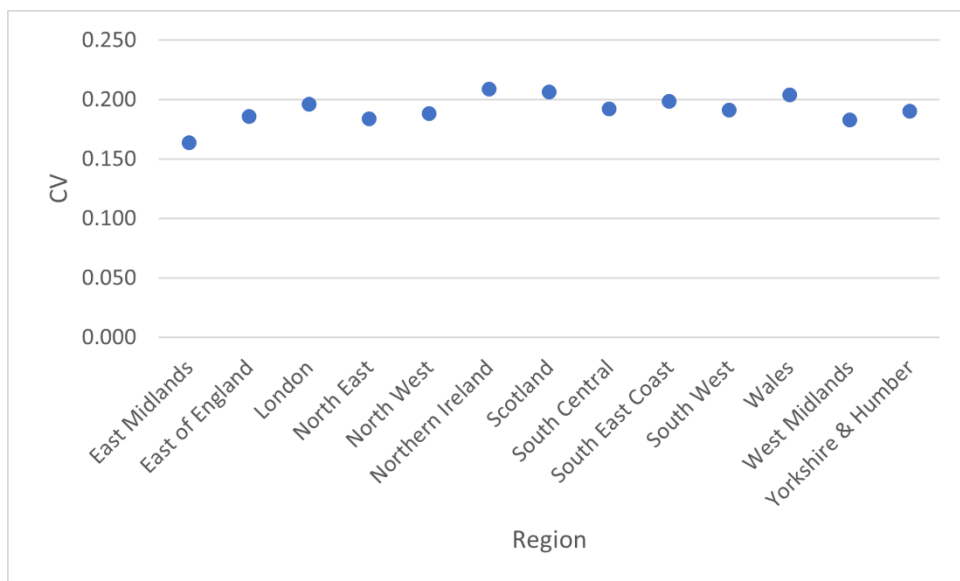


Figure 16.1 CVT by region

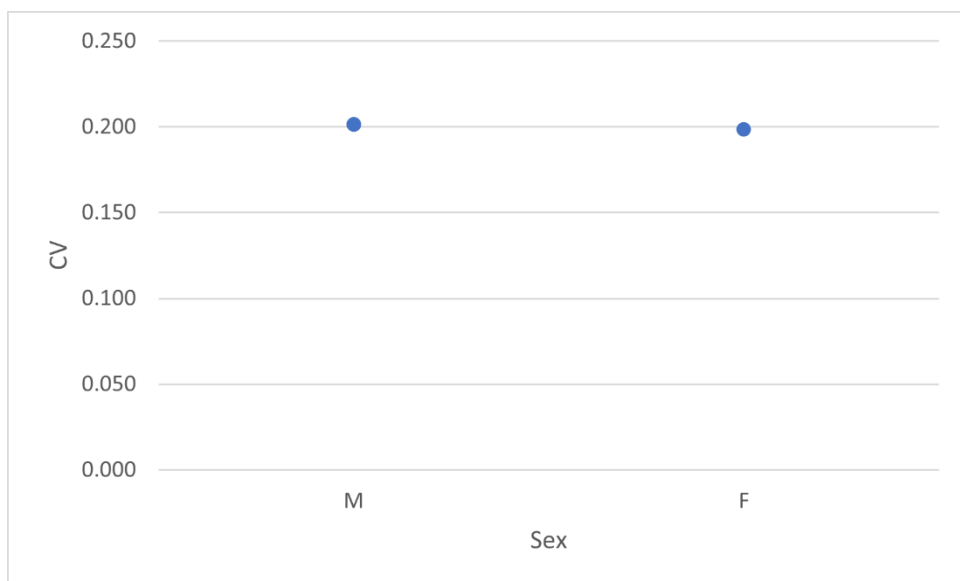


Figure 16.2 CVT by sex

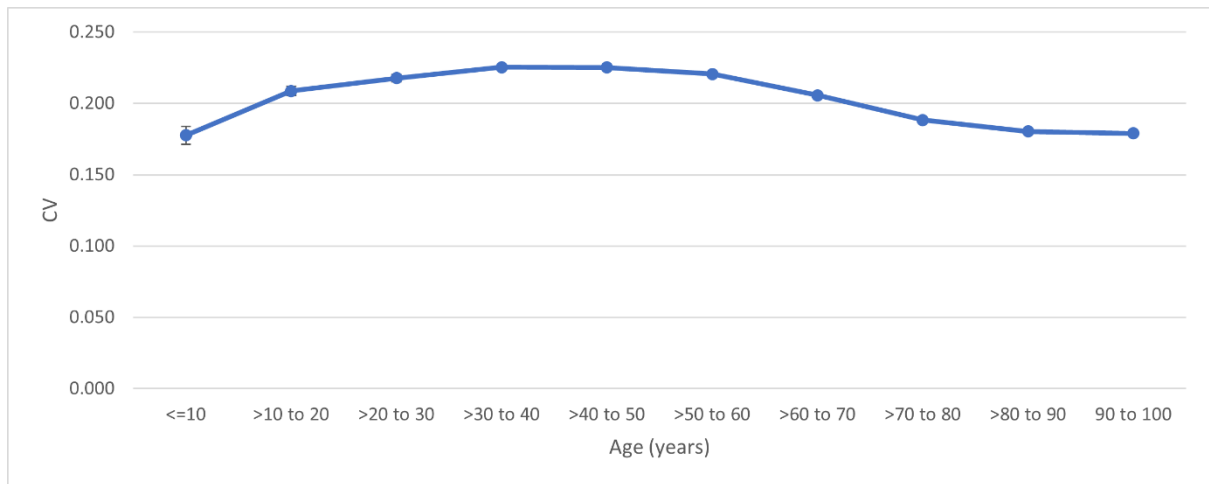


Figure 16.3 CVT by age

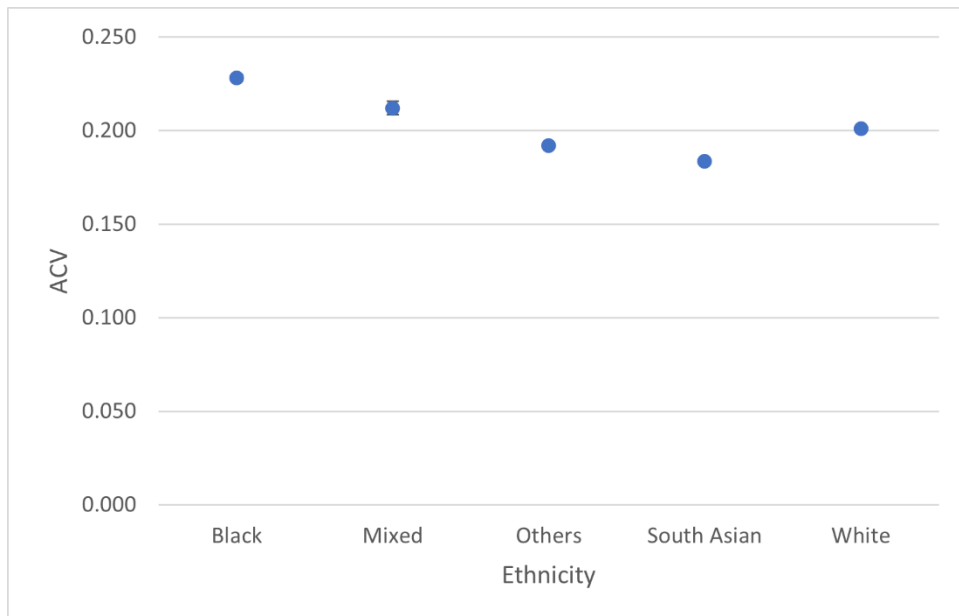


Figure 16.4 CVT by ethnicity

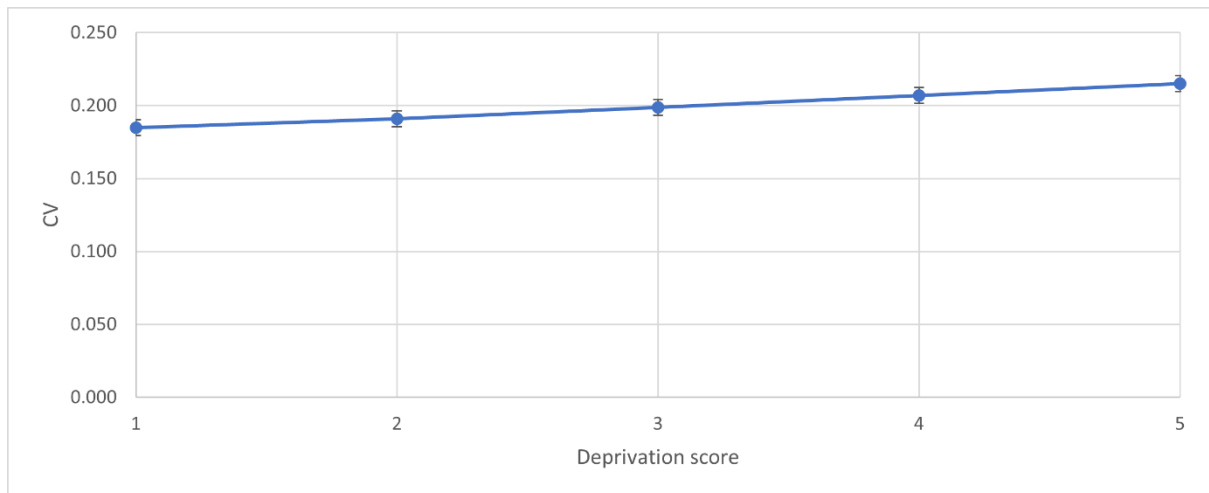


Figure 16.5 CVT by Townsend deprivation score

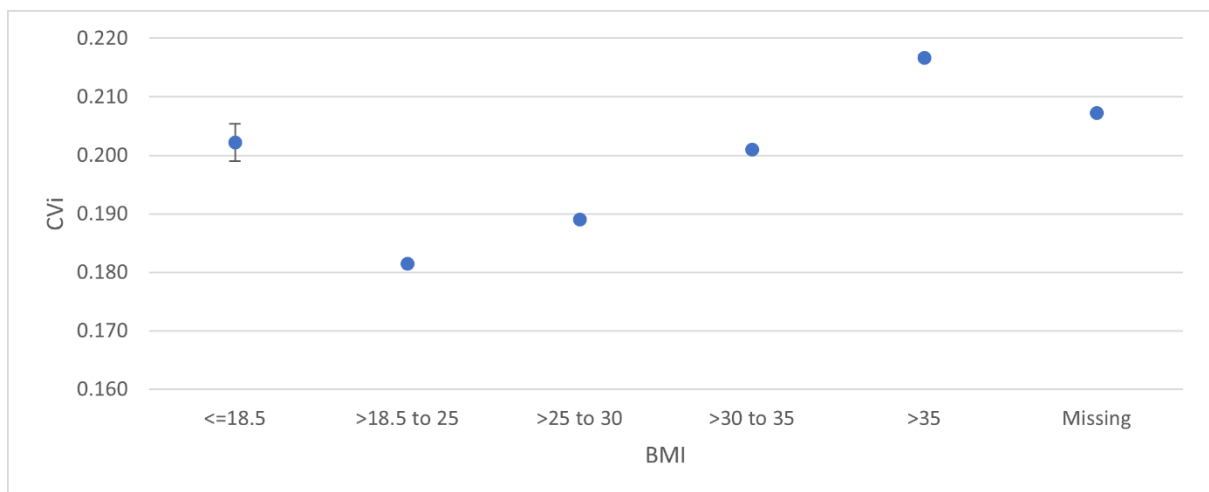


Figure 16.6 CVT by BMI

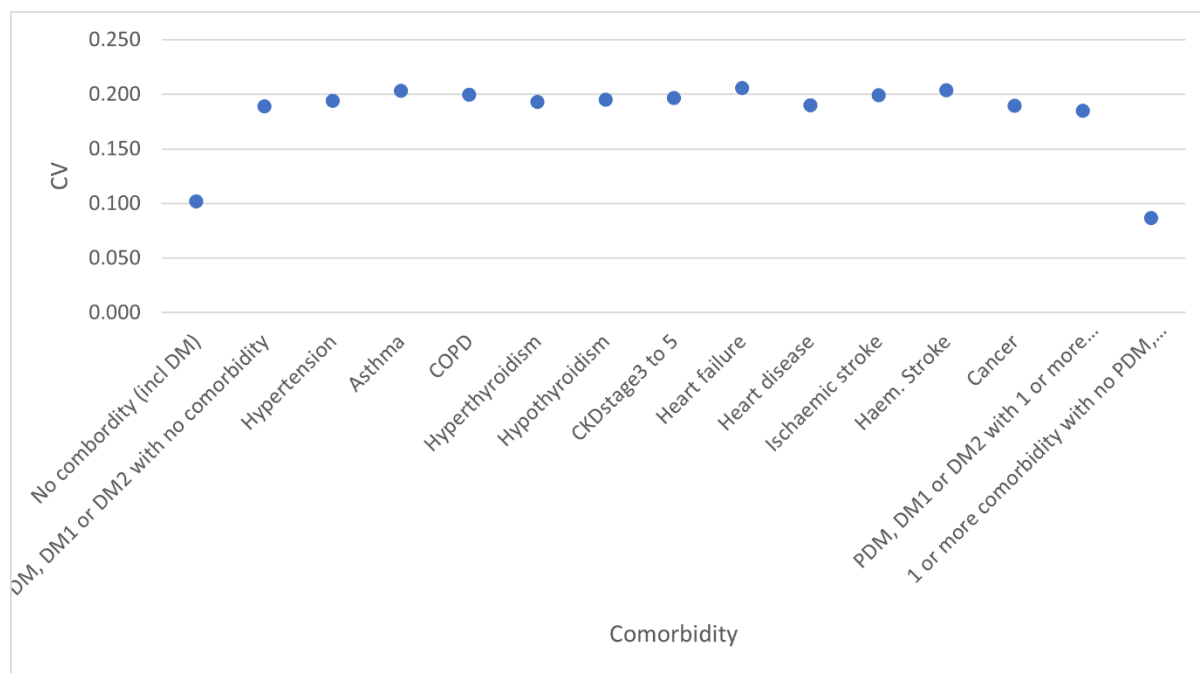


Figure 16.7 CVT by comorbidities

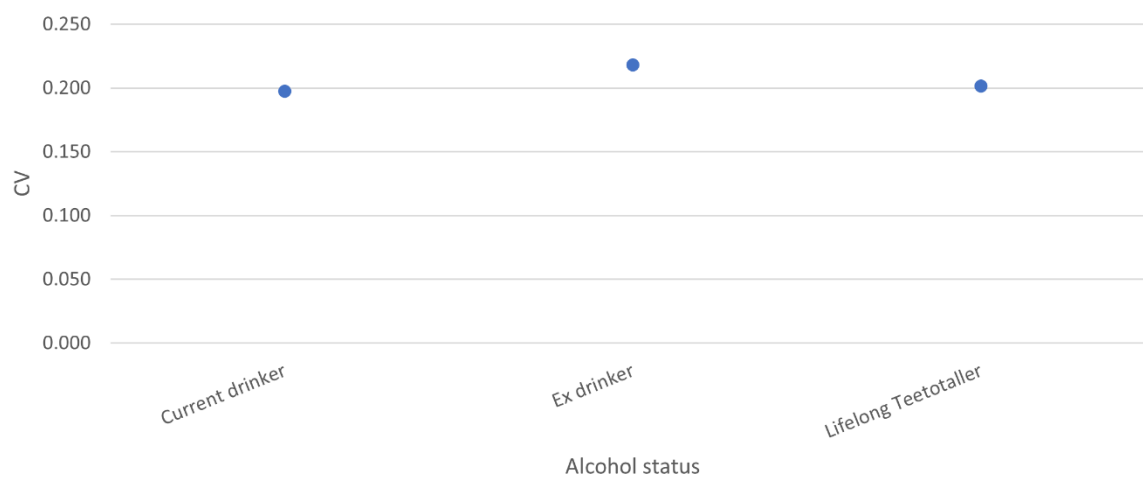


Figure 16.8 CVT by alcohol consumption

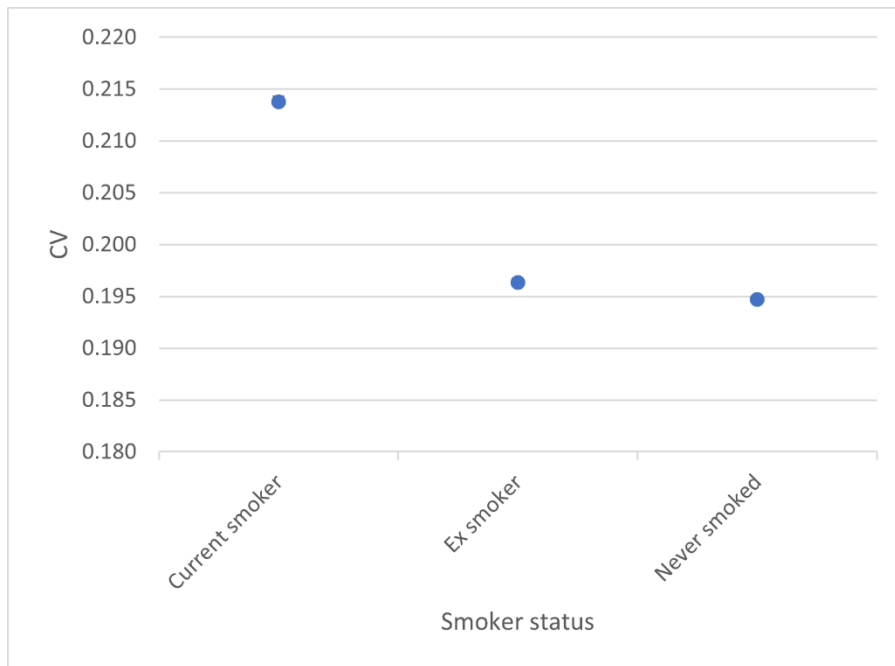


Figure 16.9 CVT by smoker status

16.2 Appendix 6.2 – sensitivity analyses

Table 16.2 Sensitivity analyses

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
Year	2010	7,300	1.2	64.92	0.158	0.157	0.160
	2011	8,963	1.5	66.31	0.162	0.161	0.164
	2012	10,972	1.9	67.53	0.159	0.158	0.160
	2013	13,351	2.3	66.48	0.161	0.160	0.162
	2014	16,725	2.8	65.27	0.154	0.153	0.155
	2015	16,738	2.9	65.88	0.156	0.155	0.157
	2016	16,310	2.8	64.88	0.151	0.150	0.152
	2017	15,225	2.6	64.47	0.153	0.152	0.154
	2018	14,218	2.4	64.18	0.153	0.152	0.154
	2019	10,188	1.7	65.01	0.159	0.157	0.160
Number of measurements	2	310,643	N/A	39.75	0.121	0.121	0.122
	3	194,601	N/A	41.92	0.138	0.138	0.139
	4	123,417	N/A	44.25	0.153	0.153	0.153
	5	83,459	N/A	46.80	0.167	0.166	0.167
	6	61,485	N/A	49.27	0.176	0.175	0.176
	7	47,972	N/A	51.24	0.180	0.179	0.181
	8	38,751	N/A	53.18	0.184	0.183	0.185
	9	31,460	N/A	55.06	0.187	0.186	0.188
	10 to 15	118,418	N/A	57.78	0.190	0.189	0.190
	16 to 20	46,905	N/A	60.58	0.189	0.189	0.190
	21 to 25	22,951	N/A	62.60	0.189	0.188	0.189
	26 to 30	8,487	N/A	65.21	0.188	0.187	0.189
	31 to 40	3,990	N/A	67.94	0.187	0.186	0.189
	41 to 50	503	N/A	69.16	0.188	0.184	0.191
	51 to 60	63	N/A	68.23	0.207	0.197	0.217
Patient median HbA1c (%)	<=5	6,853	1.2	29.86	0.128	0.127	0.129
	>5 to 6	187,834	32.0	38.49	0.113	0.113	0.114
	>6 to 7	175,186	29.8	48.42	0.166	0.166	0.167
	>7 to 8	114,466	19.5	59.24	0.187	0.186	0.187
	>8 to <9	55,352	9.4	69.69	0.193	0.192	0.193

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
	>9 to 10	26,384	4.5	79.65	0.193	0.193	0.194
	>10 to 11	12,616	2.1	89.33	0.191	0.191	0.192
	>11 to 12	5,368	0.9	98.99	0.191	0.190	0.192
	>12 to 13	2,010	0.3	108.42	0.191	0.189	0.193
	>13 to 14	699	0.1	118.26	0.196	0.192	0.200
	>14 to 15	188	0.0	126.27	0.203	0.195	0.210
Implausible values (%)	>=3 and <20	586,251	N/A	52.21	0.213	0.213	0.213
	>=4 and < 20	585,319	N/A	52.32	0.200	0.200	0.200
	All values < 100	589,255	N/A	51.79	0.240	0.240	0.241
	>=4 and <15	584,902	N/A	52.22	0.195	0.195	0.195
	>=4 and <20	585,319	N/A	52.32	0.200	0.200	0.200
	>=4 and <25	585,319	N/A	52.33	0.200	0.200	0.201
	>=4 and <30	585,319	N/A	52.33	0.200	0.200	0.201
Days between measurements	< 30	1,263	0.22	56.24	0.233	0.227	0.238
	30 to < 60	4,505	0.77	61.59	0.243	0.240	0.245
	60 to < 90	14,849	2.53	63.25	0.222	0.221	0.223
	90 to < 120	45,042	7.67	62.91	0.205	0.204	0.205
	120 to < 150	55,321	9.42	60.85	0.201	0.201	0.202
	150 to < 180	74,857	12.75	57.39	0.191	0.190	0.191
	180 to < 210	87,166	14.85	54.51	0.178	0.177	0.178
	210 to < 240	46,658	7.95	53.24	0.187	0.186	0.187
	240 to < 270	35,549	6.06	51.17	0.186	0.185	0.186
	270 to < 300	31,410	5.35	48.82	0.181	0.181	0.182
	300 to < 330	30,924	5.27	46.91	0.171	0.170	0.172
	330 to < 365	45,761	7.80	44.74	0.152	0.152	0.153
	365 to < 455	72,382	12.33	43.04	0.144	0.144	0.145
	455 to < 545	23,230	3.96	41.35	0.138	0.137	0.139
	545 to < 635	11,790	2.01	40.63	0.133	0.131	0.134
	635 to < 730	7,026	1.20	39.86	0.129	0.128	0.131
	more than 730	8,115	1.38	39.88	0.121	0.120	0.123
Method	Linear regression	587,023	100	52.31	0.200	0.200	0.201
	Arithmetic	587,023	100	55.9	0.131	0.001*	7.966*
Unit	coded as %	344,382	N/A	7.42%	0.143	0.143	0.144
	coded as mmol/mol and transformed to %	587,023	100	6.94%	0.138	0.138	0.138
	coded as mmol/mol and not transformed	587,023	100	52.31	0.200	0.200	0.201

Mean is in mmol/mol unless otherwise stated. *=range.

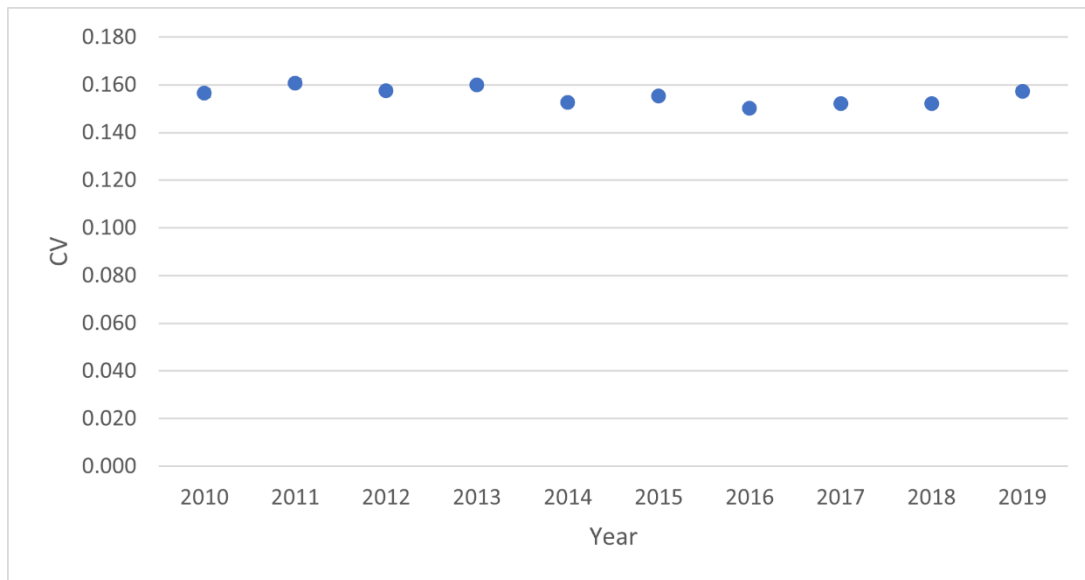


Figure 16.10 CVT by year

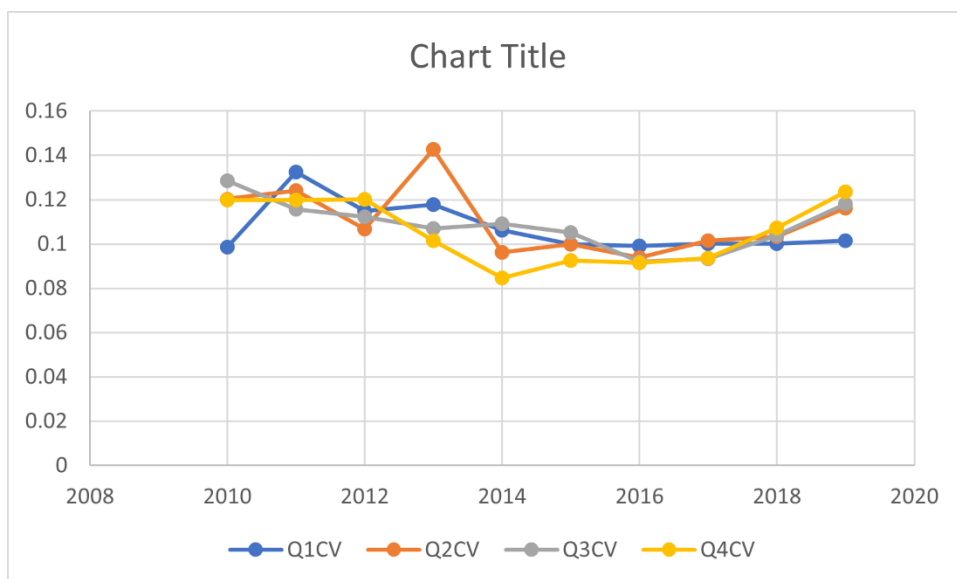


Figure 16.11 CVT by quarter

Table 16.3 CVT by quarter

Year		Q1	Q2	Q3	Q4
2010		0.099	0.120	0.128	0.120
2011		0.133	0.124	0.116	0.120
2012		0.115	0.107	0.112	0.120
2013		0.118	0.143	0.107	0.101
2014		0.106	0.096	0.109	0.085
2015		0.100	0.100	0.105	0.093
2016		0.099	0.094	0.092	0.092
2017		0.100	0.101	0.093	0.094
2018		0.100	0.103	0.104	0.107
2019		0.101	0.116	0.118	0.123

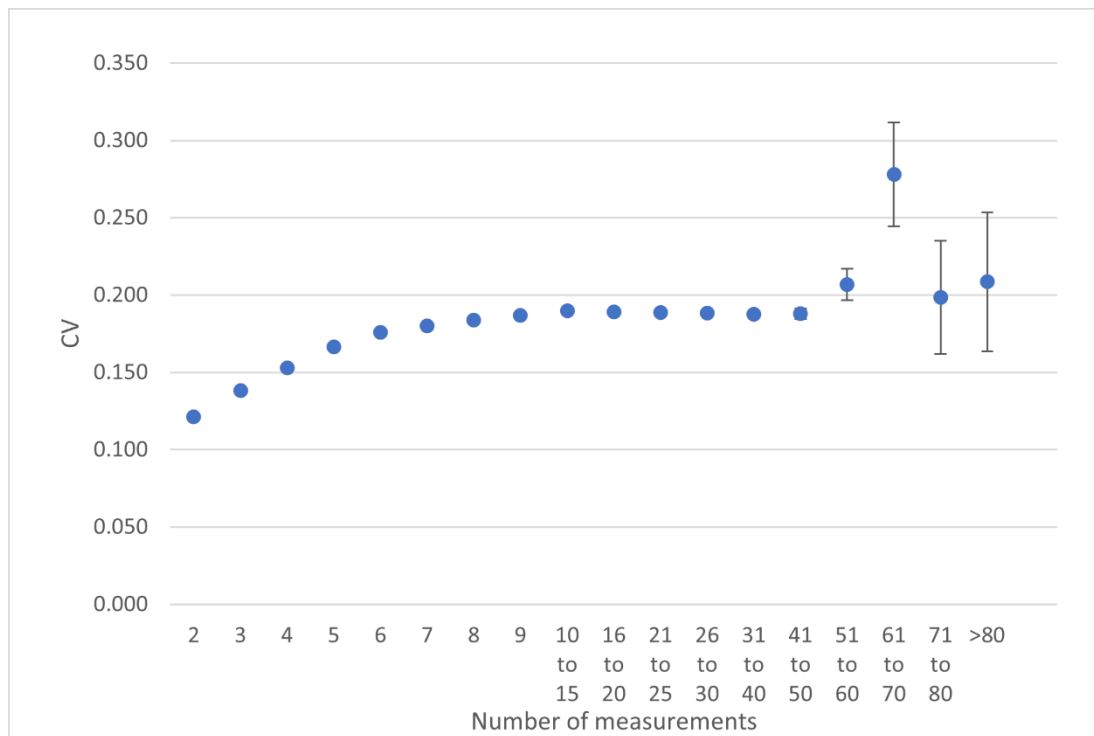


Figure 16.12 CVT by Number of measurements

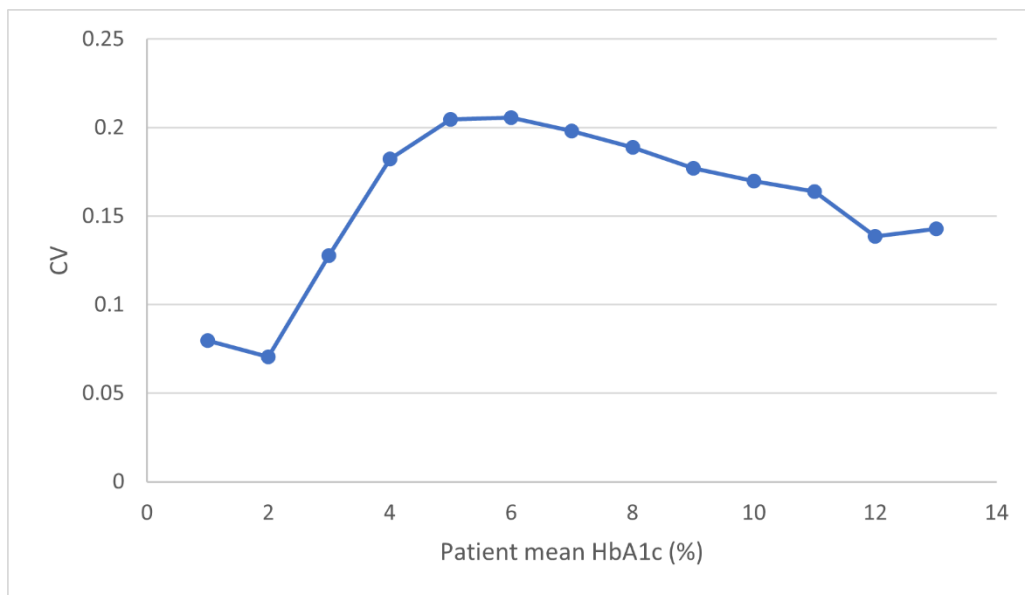


Figure 16.13 CVT by Mean

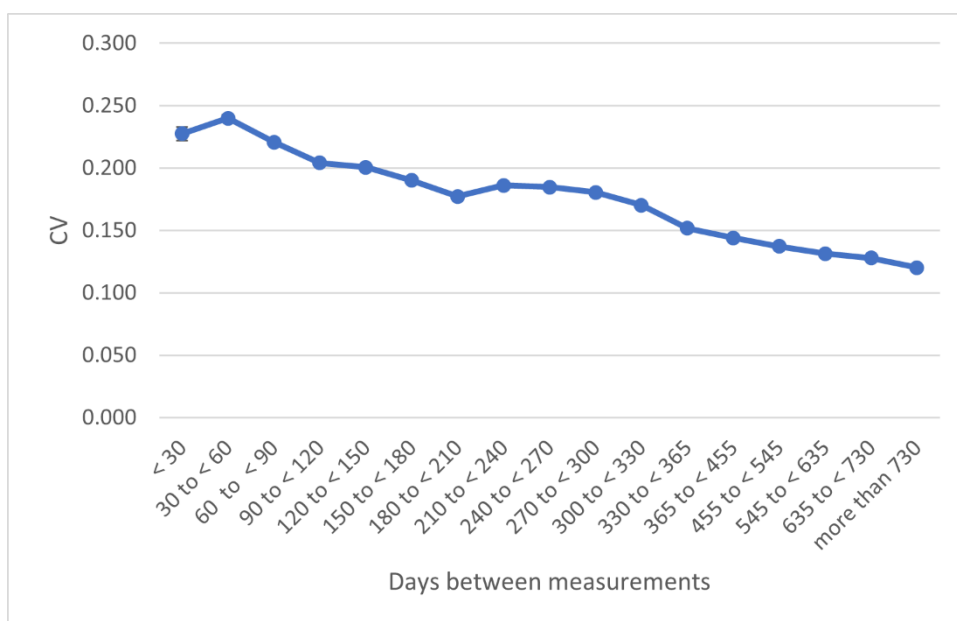


Figure 16.14 CVT by days between measurements

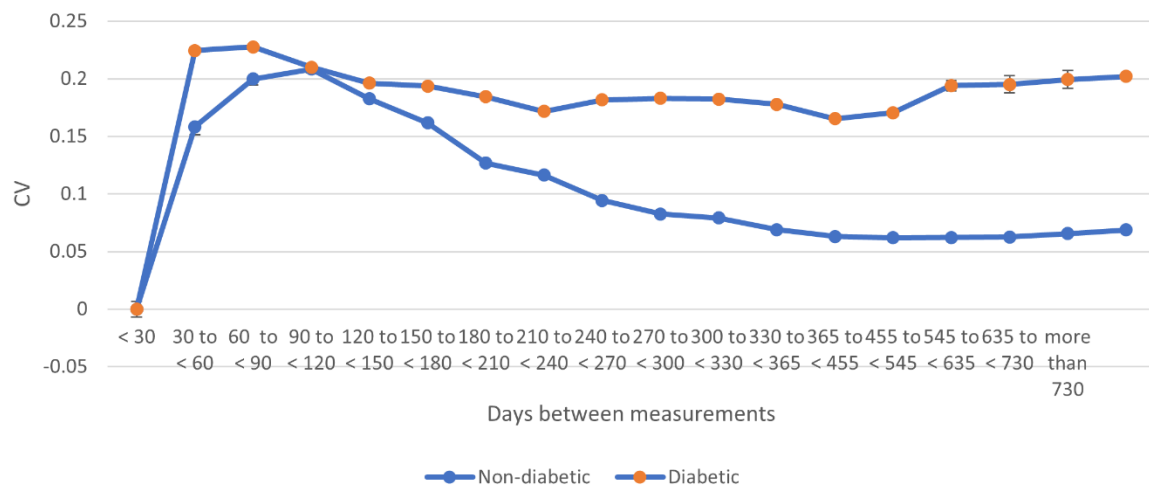


Figure 16.15 CVT by Days between measurements stratified by diabetic status

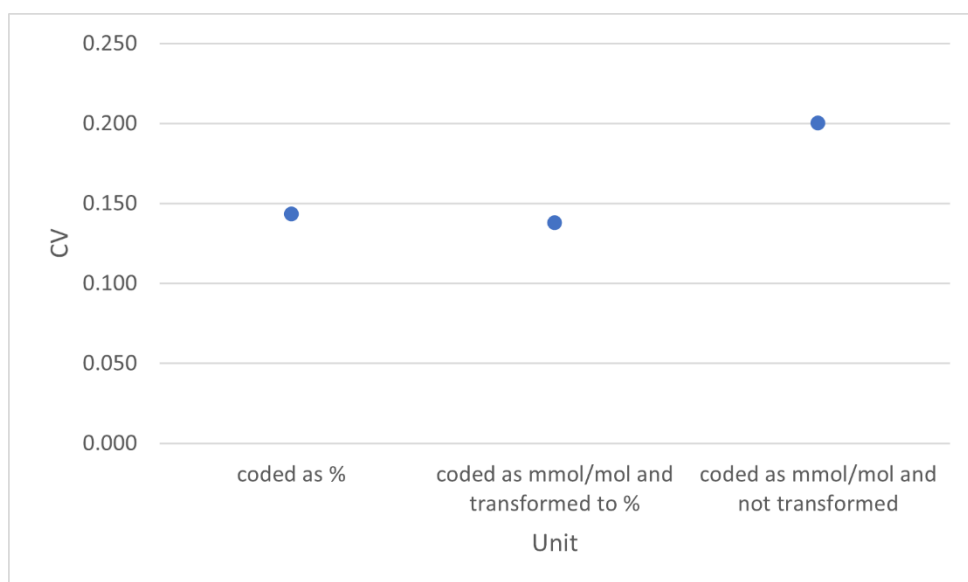


Figure 16.16 CVT by unit of measurement

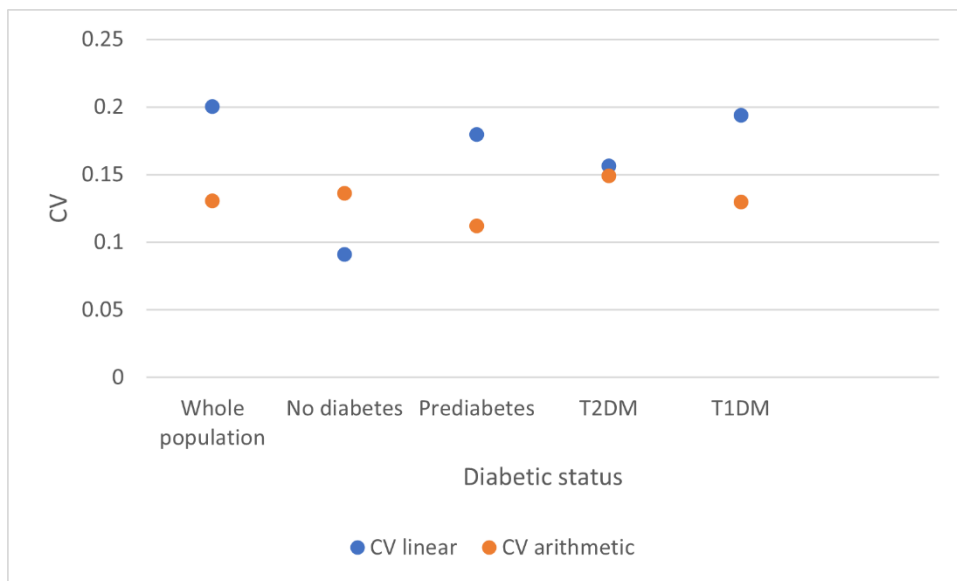


Figure 16.17 CVT by arithmetic calculation methods

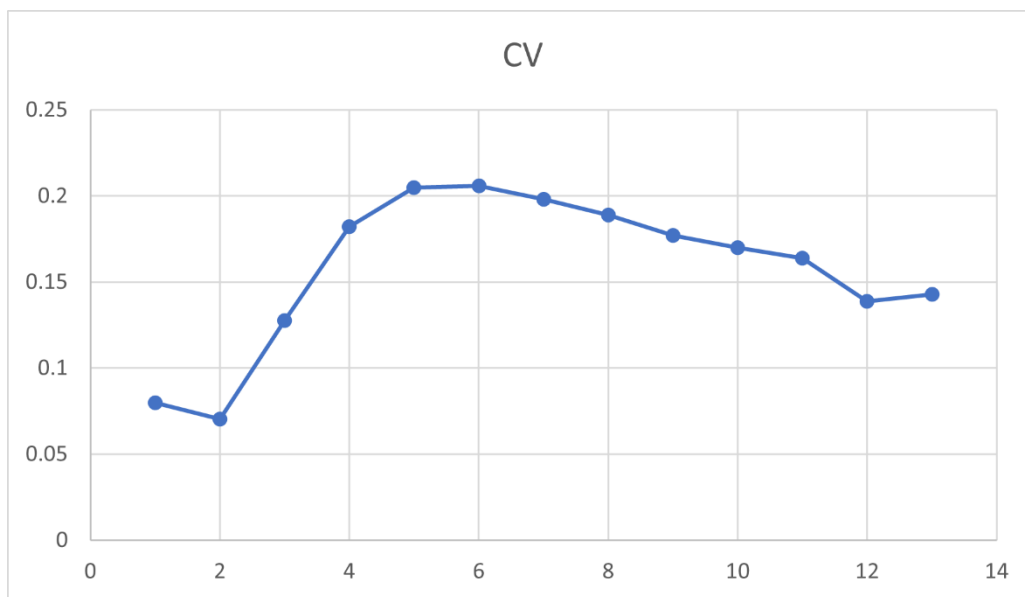


Figure 16.18 CVT by mean

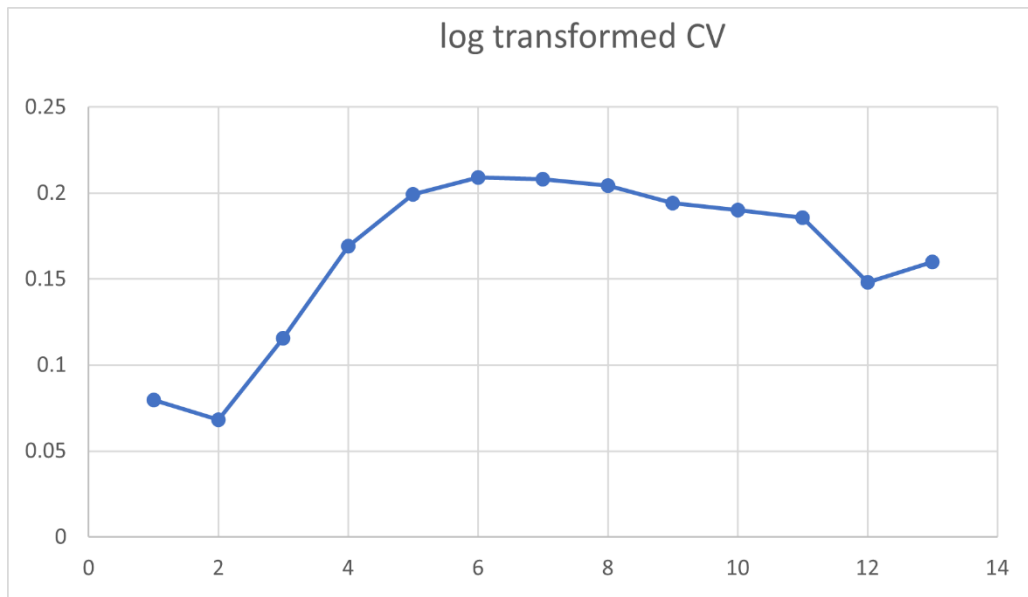


Figure 16.19 CVT by mean (log-transformed data)

17 Appendix 7 CRP cohort study appendices

17.1 Appendix 7.1 - Subgroup analysis

Table 17.1 Other subgroup analyses

		N	%	Median (mg/mL)	Log adj CV (95%CI)	95% CI
Comorbidities	No comorbidity	125,203	26.5	4.0	1.124	1.122 - 1.126
	Hypertension	160,359	33.9	5.8	1.167	1.165 - 1.169
	Asthma	93,668	19.8	5.0	1.121	1.118 - 1.124
	COPD	35,404	7.5	8.0	1.245	1.240 - 1.250
	Hyperthyroidism	10,573	2.2	5.0	1.096	1.088 - 1.103
	Hypothyroidism	50,917	10.8	5.0	1.076	1.073 - 1.079
	CKDstage3 to 5	54,147	11.5	6.0	1.226	1.222 - 1.230
	Heart failure	12,728	2.7	8.0	1.239	1.230 - 1.247
	Heart disease	52,131	11.0	5.2	1.239	1.235 - 1.244
	Ischaemic stroke	6,777	1.4	6.0	1.303	1.290 - 1.316
	Haem. Stroke	2,451	0.5	6.0	1.232	1.213 - 1.252
	Cancer	38,639	8.2	5.7	1.258	1.252 - 1.263
	PreDM or DM	5,960	1.3	5.0	1.061	1.051 - 1.071
BMI	≤18.5	8,630	1.8	4.4	1.286	1.275 - 1.297
	>18.5 to 25	136,598	28.9	4.7	1.214	1.211 - 1.216
	>25 to 30	129,716	27.4	5.0	1.139	1.137 - 1.141
	>30 to 35	60,409	12.8	6.0	1.035	1.032 - 1.038
	>35	137,240	29.0	5.0	1.131	1.129 - 1.133
Region	East Midlands	2,734	0.6	6.0	1.006	0.991 - 1.021
	East of England	17,097	3.6	6.0	1.015	1.010 - 1.021
	London	30,912	6.5	4.2	1.191	1.185 - 1.197
	North East	2,505	0.5	7.0	0.971	0.960 - 0.982
	North West	42,535	9.0	4.0	1.142	1.138 - 1.145
	Northern Ireland	58,609	12.4	4.0	1.272	1.269 - 1.276

	Scotland	66,526	14.1	6.0	1.081	1.078 - 1.085
	South Central	54,242	11.5	5.0	1.247	1.243 - 1.250
	South East Coast	46,868	9.9	5.0	1.205	1.201 - 1.210
	South West	40,817	8.6	5.0	1.155	1.151 - 1.159
	Wales	82,138	17.4	5.0	1.044	1.041 - 1.046
	West Midlands	24,523	5.2	5.0	0.996	0.991 - 1.000
	Yorkshire & Humber	3,087	0.7	5.3	1.128	1.113 - 1.143
Alcohol consumption	Current drinker	272,933	57.8	5.0	1.148	1.146 - 1.149
	Ex drinker	6,649	1.4	5.0	1.134	1.124 - 1.145
	Teetotal	85,881	18.2	5.0	1.096	1.093 - 1.098
Smoker Status	Current Smoker	99,383	21.03	5.0	1.118	1.116 - 1.121
	Ex-smoker	87,664	18.55	5.0	1.159	1.156 - 1.162
	Never smoked	222,385	47.06	5.0	1.134	1.132 - 1.136

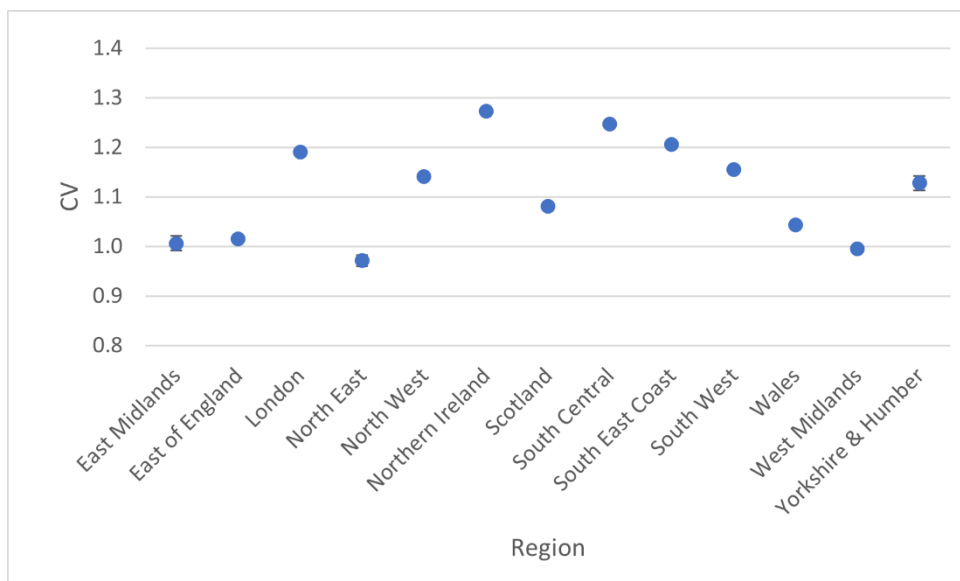


Figure 17.1 CVT by region

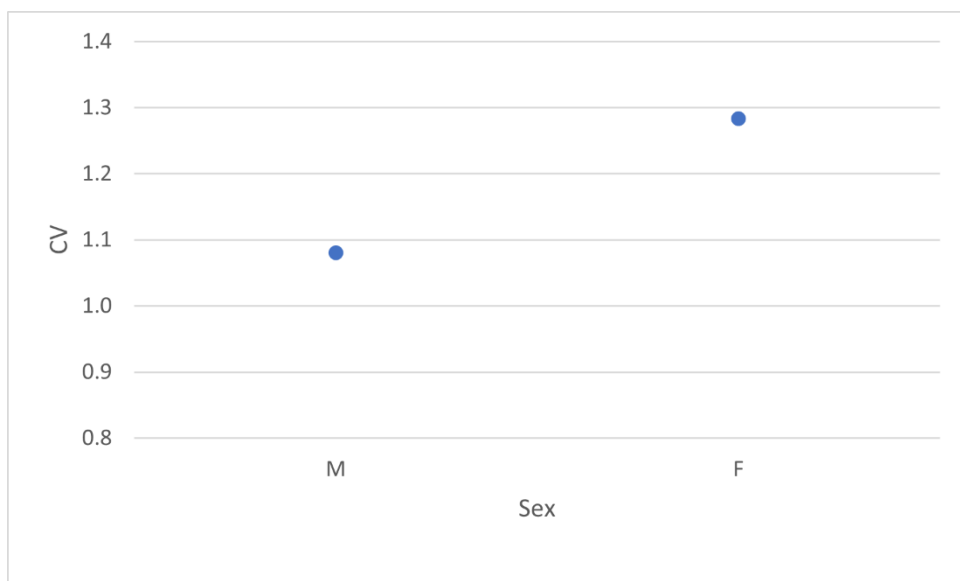


Figure 17.2 CVT by sex

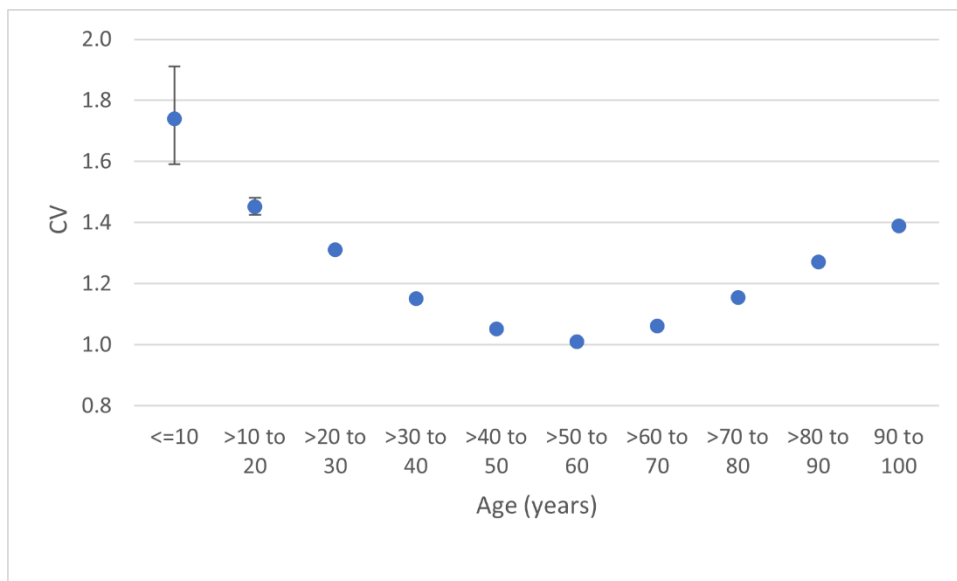


Figure 17.3 CVT by age

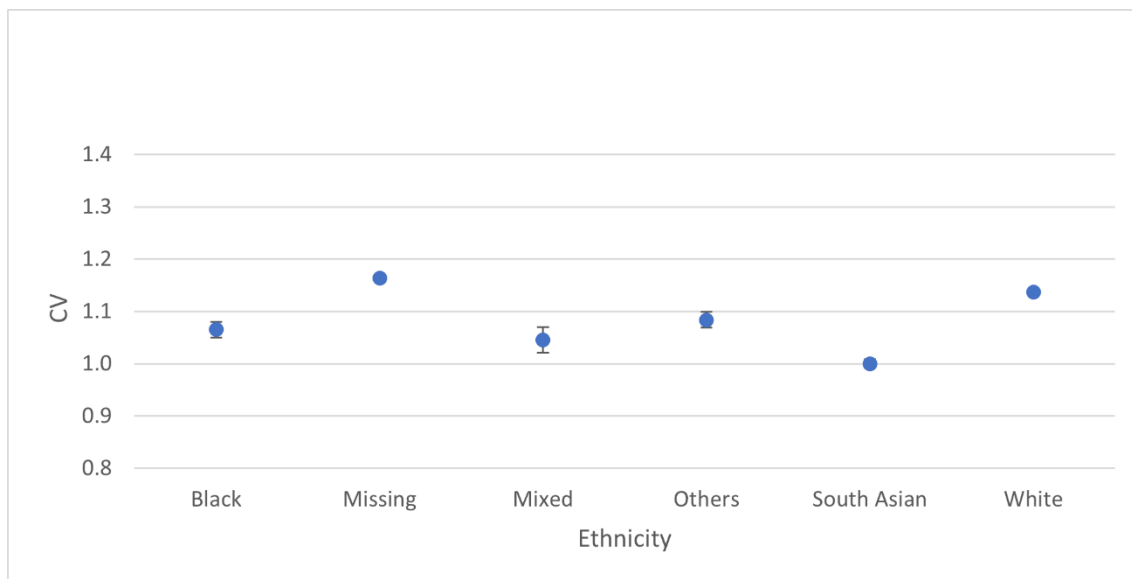


Figure 17.4 CVT by ethnicity

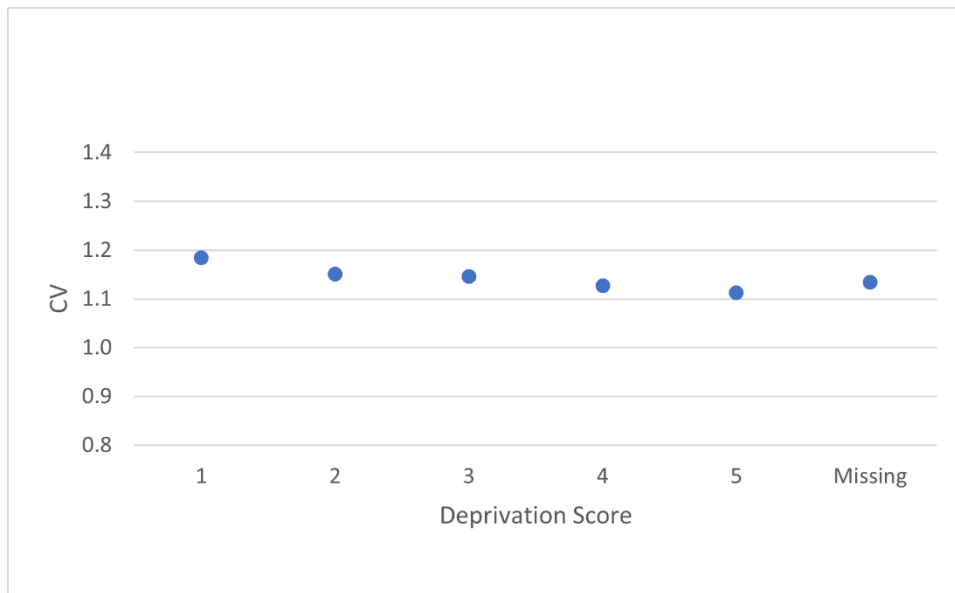


Figure 17.5 CVT by Townsend deprivation score

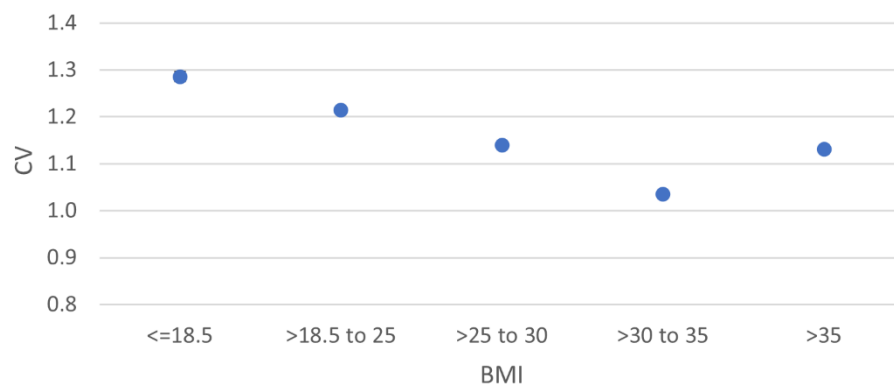


Figure 17.6 CVT by BMI

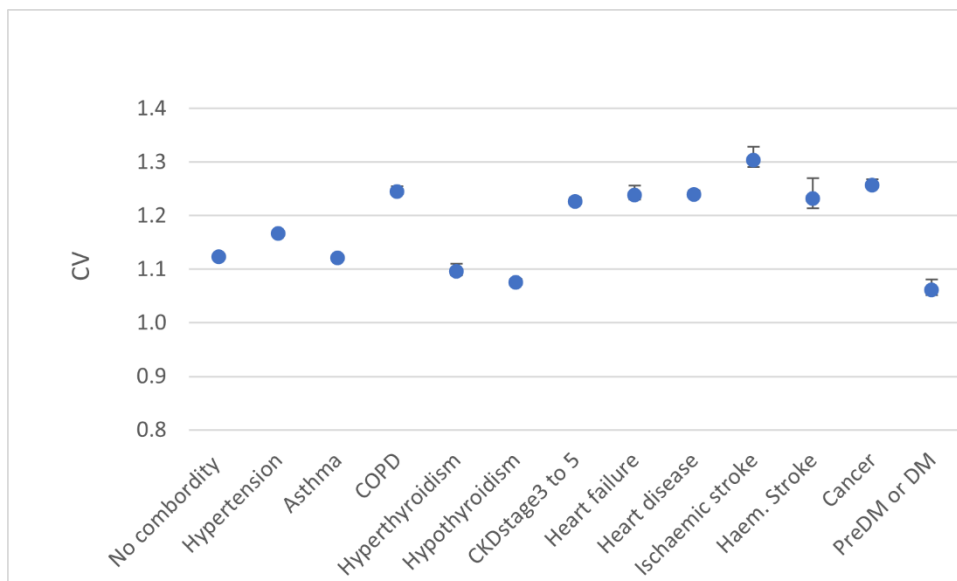


Figure 17.7 CVT by comorbidities

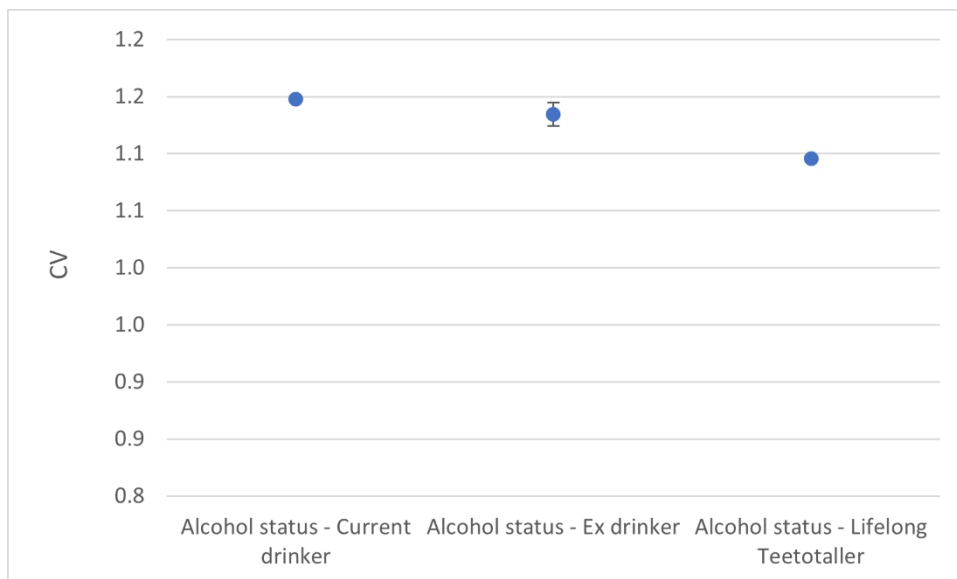


Figure 17.8 CVT by alcohol consumption

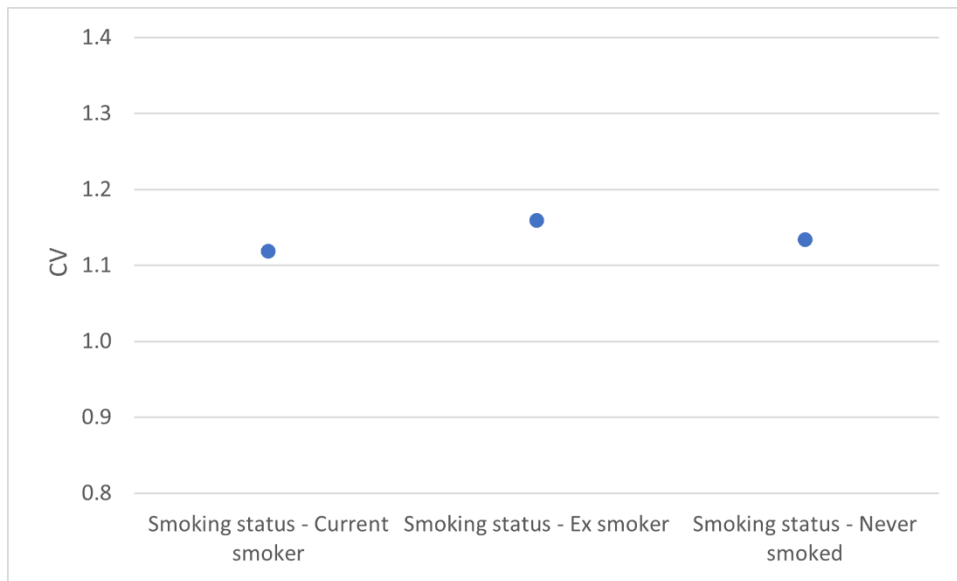


Figure 17.9 CVT by smoker status

17.2 Appendix 7.2 – sensitivity analyses

Table 17.2 Sensitivity analyses

*=mean

		N	%	Median (mg/L)	CV (95%CI)	95% CIs
Year	2010	14,663	3.1	6.0	0.831	0.826 - 0.836
	2011	17,219	3.6	5.9	0.883	0.878 - 0.888
	2012	19,028	4.0	5.0	0.884	0.880 - 0.889
	2013	20,329	4.3	5.0	0.920	0.915 - 0.925
	2014	20,667	4.4	5.0	0.925	0.910 - 0.919
	2015	19,257	4.1	5.0	0.899	0.894 - 0.904
	2016	15,867	3.4	5.0	0.938	0.933 - 0.944
	2017	14,397	3.0	5.0	0.907	0.901 - 0.913
	2018	13,920	2.9	5.0	0.958	0.952 - 0.965
	2019	13,533	2.9	6.0	0.937	0.931 - 0.944
Number of measurements	2	422,639	N/A	5.0	1.164	1.104 - 1.229
	3	248,438	N/A	4.9	1.191	1.162 - 1.221
	4	146,997	N/A	5.0	1.265	1.261 - 1.269
	5	90,749	N/A	5.0	1.246	1.241 - 1.250
	6	59,023	N/A	5.0	1.231	1.226 - 1.236
	7	39,132	N/A	5.0	1.215	1.210 - 1.220
	8	26,919	N/A	5.0	1.218	1.212 - 1.224
	9	19,107	N/A	5.0	1.207	1.201 - 1.214
	10 to 15	48,151	N/A	5.0	1.197	1.193 - 1.200
	16 to 20	13,190	N/A	5.0	1.127	1.122 - 1.132
	21 to 25	6,925	N/A	5.0	1.089	1.084 - 1.095
	26 to 30	4,547	N/A	5.1	1.066	1.060 - 1.072
	31 to 40	5,674	N/A	5.0	1.022	1.018 - 1.027
	41 to 50	3,578	N/A	5.0	0.997	0.992 - 1.002
	51 to 60	2,357	N/A	5.0	1.002	0.997 - 1.007
	61 to 70	1,674	N/A	5.0	0.983	0.978 - 0.989
	71 to 80	1,225	N/A	5.5	0.976	0.970 - 0.983
	81 to 90	910	N/A	5.0	0.960	0.953 - 0.966
	91 to 100	618	N/A	5.3	0.935	0.927 - 0.942
	>100	1,817	N/A	6.0	1.050	1.046 - 1.055
Patient mean CRP (mg/L)	0 to <1	9,095	1.9	0.70*	0.692	0.687- 0.697
	1 to < 2	36,942	7.8	1.50*	0.661	0.659- 0.664

		N	%	Median (mg/L)	CV (95%CI)	95% CIs
	2 to <3	40,887	8.7	2.47*	0.824	0.822- 0.827
	3 to < 4	41,918	8.9	3.47*	0.848	0.845- 0.850
	4 to <5	38,630	8.2	4.45*	0.831	0.829- 0.834
	5 to <6	32,432	6.9	5.45*	0.839	0.836-0.842
	6 to < 7	26,605	5.6	6.45*	0.903	0.899- 0.906
	7 to < 8	22,505	4.8	7.45*	0.954	0.951- 0.958
	8 to < 9	19,108	4.0	8.45*	1.008	1.004- 1.013
	9 to < 10	16,741	3.5	9.45*	1.050	1.045- 1.055
	10 to <20	93,129	19.7	14.12*	1.224	1.221-1.227
	20 to < 30	39,071	8.3	24.25*	1.563	1.556- 1.569
	30 to < 40	20,707	4.4	34.39*	1.837	1.826- 1.849
	40 to < 50	11,910	2.5	44.49*	2.115	2.095- 2.136
	50 to < 60	7,614	1.6	54.51*	2.301	2.271-2.331
	60 to < 70	4,869	1.0	64.57*	2.541	2.497- 2.587
	70 to < 80	3,377	0.7	74.66*	2.658	2.600- 2.719
	80 to < 90	2,123	0.4	84.55*	2.688	2.612-2.767
	90 to < 100	1,440	0.3	94.45*	2.673	2.582- 2.769
	100 to < 200	3,363	0.7	124.71*	2.733	2.669-2.800
	200 to < 300	119	0.07	229.16*	1.848	1.666- 2.064
Days between measurements	0 to < 4	997	0.2	34.0	1.728	1.672 - 1.788
	4 to < 7	1,440	0.3	22.0	1.834	1.779 - 1.892
	7 to < 14	6,895	1.5	12.0	1.458	1.443 - 1.473
	14 to < 21	9,077	1.9	9.0	1.249	1.241 - 1.257
	21 to < 28	8,278	1.8	8.9	1.266	1.257 - 1.275
	28 to < 35	19,505	4.1	6.0	1.011	1.008 - 1.013
	35 to < 42	10,745	2.3	6.0	1.099	1.094 - 1.104
	42 to < 49	9,166	1.9	6.0	1.124	1.118 - 1.130
	49 to 60	14,001	3.0	6.0	1.139	1.134 - 1.144
	60 to < 90	28,335	6.0	6.0	1.184	1.180 - 1.189
	90 to < 120	25,069	5.3	5.0	1.228	1.222 - 1.234
	120 to < 150	21,128	4.5	5.0	1.242	1.236 - 1.249
	150 to < 180	20,693	4.4	5.0	1.219	1.212 - 1.226
	180 to < 270	58,220	12.3	5.0	1.191	1.187 - 1.195
	270 to < 365	54,605	11.6	4.0	1.146	1.142 - 1.150
	365 to < 455	41,779	8.8	4.0	1.122	1.117 - 1.127
	455 to < 545	32,148	6.8	4.0	1.130	1.124 - 1.136

		N	%	Median (mg/L)	CV (95%CI)	95% CIs
	545 to < 635	25,154	5.3	4.0	1.144	1.137 – 1.151
	635 to < 730	20,876	4.4	4.0	1.130	1.122 – 1.138
	>730	64,362	13.6	5.0	1.153	1.149 – 1.158
Method	Linear regression (log adjusted data)	472,593	100	5.0	1.148	1.147 - 1.149
	Linear regression (non log-adjusted data)	472,593	100	5.0	1.842	1.835 – 1.849
	Arithmetic method	472,593	100	5.0	0.998	0.005 - 5.869 (range)

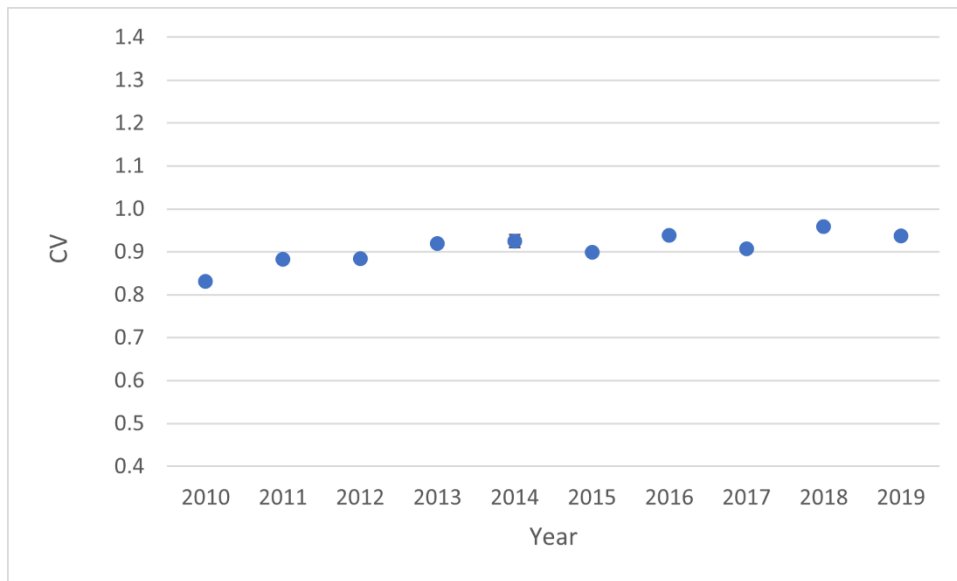


Figure 17.10 CVT by year

Table 17.3 CVT by quarter

Year	Q1	Q2	Q3	Q4
2010	0.844 (0.823-0.867)	0.824 (0.803-0.845)	0.815 (0.796-0.835)	0.837 (0.816-0.858)
2011	0.870 (0.850-0.891)	0.853 (0.833-0.873)	0.832 (0.814-0.850)	0.869 (0.849-0.890)
2012	0.913 (0.893-0.934)	0.868 (0.849-0.887)	0.839 (0.821-0.858)	0.840 (0.822-0.859)
2013	0.918 (0.897-0.939)	0.916 (0.896-0.936)	0.916 (0.897-0.937)	0.865 (0.845-0.884)
2014	0.932 (0.912-0.954)	0.932 (0.911-0.953)	0.905 (0.886-0.925)	0.920 (0.899-0.942)
2015	0.942 (0.920-0.964)	0.831 (0.812-0.849)	0.856 (0.837-0.875)	0.869 (0.848-0.890)
2016	0.986 (0.961-1.011)	0.957 (0.933-0.981)	0.921 (0.898-0.945)	0.945 (0.919-0.973)
2017	0.934 (0.910-0.959)	0.916 (0.892-0.942)	0.888 (0.864-0.913)	0.944 (0.916-0.973)
2018	1.035 (1.004-1.066)	0.945 (0.918-0.973)	0.938 (0.911-0.966)	0.965 (0.936-0.994)
2019	0.946 (0.919-0.975)	0.926 (0.898-0.954)	0.934 (0.908-0.962)	0.901 (0.876-0.928)

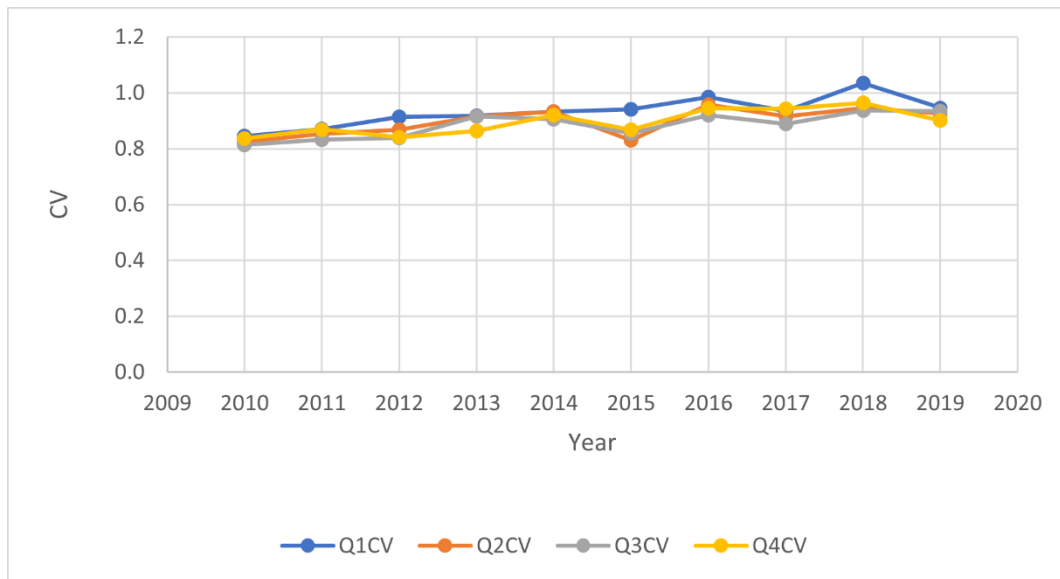


Figure 17.11 CVT by quarter

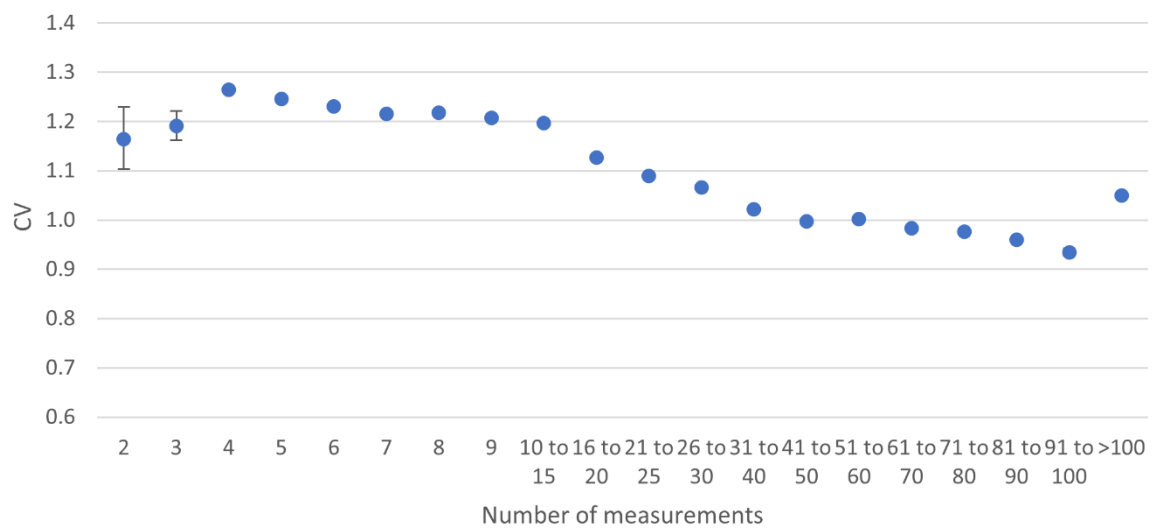


Figure 17.12 CVT by Number of measurements

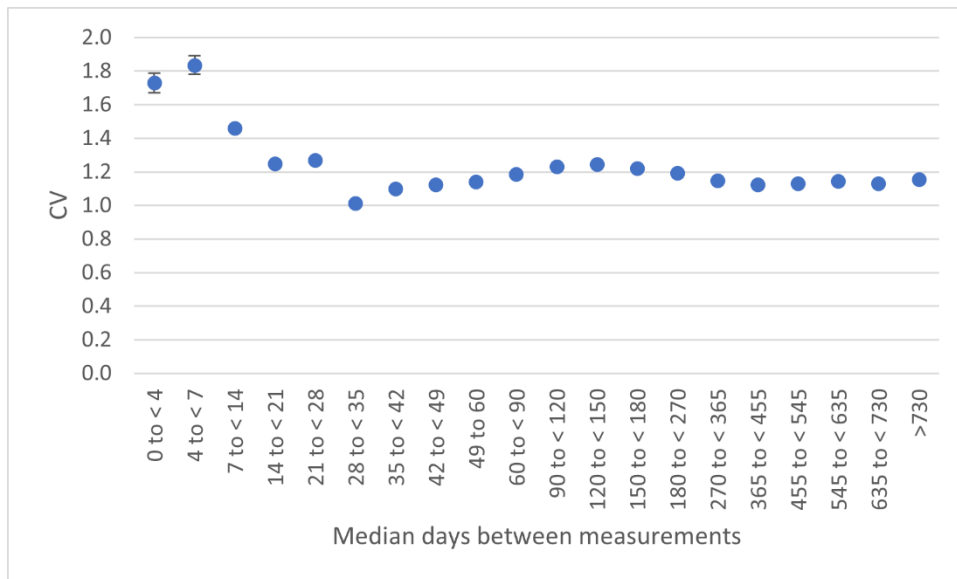


Figure 17.13 CVT by days between measurements

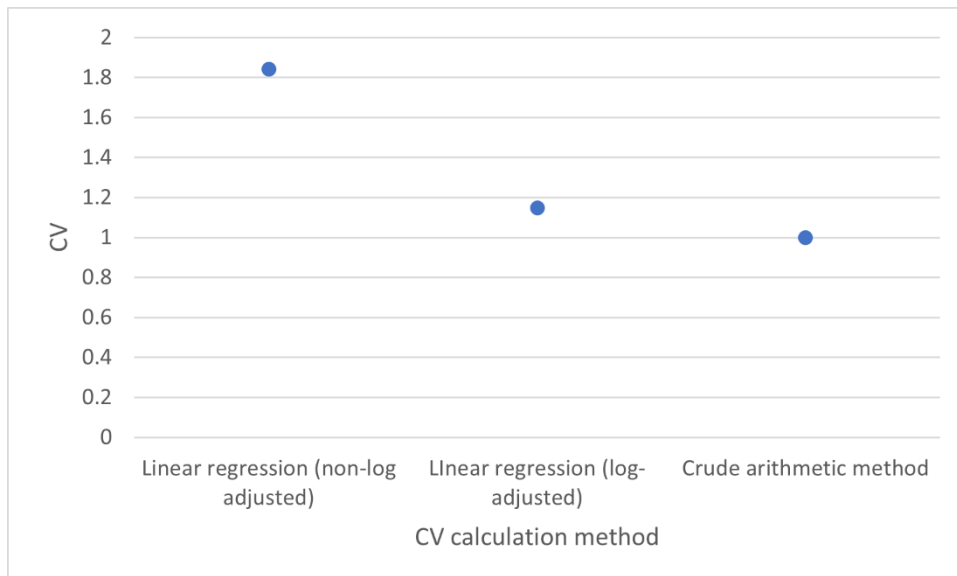


Figure 17.14 CVT by arithmetic calculation methods (using non-log adjusted data) and linear regression (using non-log adjusted data)

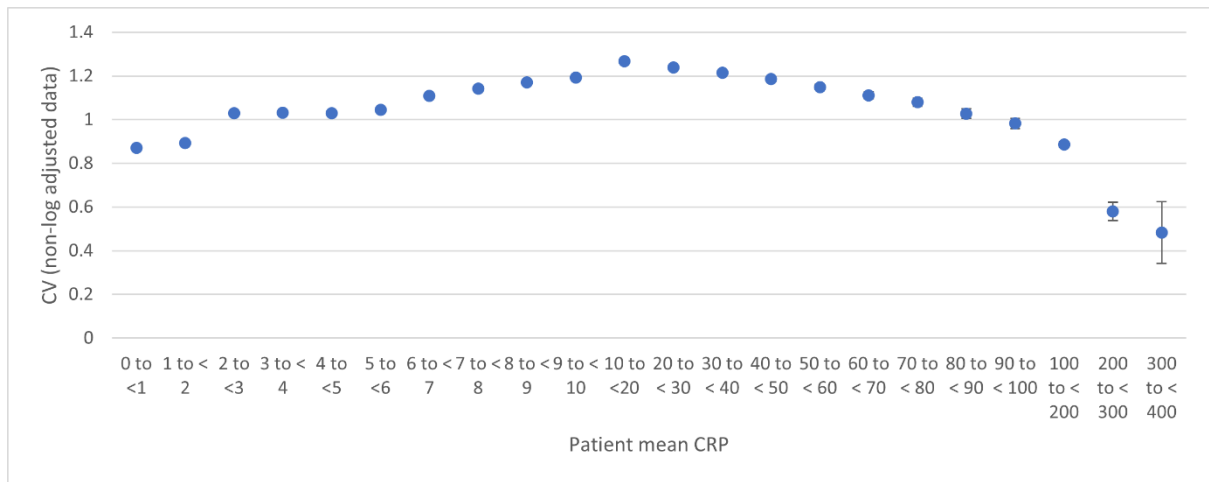


Figure 17.15 CVT by patient mean (non-log adjusted data)

18 Appendix 8 Spirometry cohort study appendices

18.1 Appendix 8.1 - Subgroup analysis

Table 18.1 Subgroup analyses for FEV1

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
Comorbidities	No comorbidity	229	5.2	2.73	0.135	0.127	0.143
	Hypertension	1,417	32.1	1.83	0.243	0.236	0.249
	Asthma	1,805	40.9	2.05	0.216	0.210	0.221
	COPD	1,514	34.3	1.72	0.291	0.284	0.299
	Hyperthyroidism	78	1.8	1.67	0.218	0.194	0.243
	Hypothyroidism	280	6.3	1.77	0.209	0.196	0.221
	CKDstage3 to 5	245	5.6	1.72	0.237	0.222	0.251
	Heart failure	112	2.5	1.72	0.291	0.262	0.319
	Ischaemic heart disease	563	12.8	1.83	0.234	0.224	0.244
	Ischaemic stroke	50	1.1	1.76	0.241	0.207	0.275
	Haem. Stroke	20	0.5	x	x	x	x
	Cancer	236	5.3	1.72	0.249	0.232	0.265
	DM or Pre DM	19	0.4	x	x	x	x
Deprivation score	1	747	16.9	2.11	0.202	0.194	0.210
	2	786	17.8	2.04	0.212	0.204	0.220
	3	904	20.5	2.04	0.236	0.227	0.244
	4	923	20.9	1.99	0.236	0.228	0.244
	5	598	13.6	1.96	0.239	0.228	0.249
	Missing	454	10.3	1.99	0.219	0.208	0.230
Region	East Midlands	72	1.6	2.12	0.168	0.144	0.191
	East of England	283	6.4	2.02	0.188	0.176	0.200
	London	390	8.8	1.99	0.199	0.188	0.210
	North East	120	2.7	1.83	0.251	0.226	0.277
	North West	552	12.5	1.96	0.244	0.233	0.255
	Northern Ireland	168	3.8	2.16	0.217	0.200	0.233
	Scotland	451	10.2	2.04	0.225	0.214	0.236
	South Central	371	8.4	2.11	0.212	0.200	0.223
	South East Coast	372	8.4	2.06	0.216	0.205	0.228
	South West	264	6.0	1.99	0.231	0.217	0.246
	Wales	482	10.9	1.99	0.213	0.203	0.223
	West Midlands	692	15.7	2.04	0.255	0.245	0.265
	Yorkshire & Humber	195	4.4	2.08	0.207	0.191	0.224
Alcohol consumption	Current drinker	2,775	62.9	2.05	0.227	0.222	0.231
	Ex-drinker	71	1.6	2.09	0.253	0.223	0.282
	Teetotal	610	13.8	1.83	0.245	0.235	0.256

Table 18.2 Subgroup analysis for FVC

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
Comorbidities	No comorbidity	201	5.6	3.59	0.120	0.112	0.127
	Hypertension	1,131	31.7	2.79	0.165	0.161	0.170
	Asthma	1,422	39.9	3.08	0.149	0.145	0.153
	COPD	1,087	30.5	2.80	0.175	0.169	0.180
	Hyperthyroidism	61	1.7	2.55	0.165	0.146	0.185
	Hypothyroidism	245	6.9	2.62	0.175	0.165	0.186
	CKDstage3 to 5	211	5.9	2.59	0.177	0.166	0.188
	Heart failure	90	2.5	2.75	0.217	0.195	0.240
	Ischaemic heart disease	405	11.4	2.80	0.183	0.174	0.191
	Ischaemic stroke	39	1.1	x	x	x	x
	Haem. Stroke	17	0.5	x	x	x	x
	Cancer	198	5.6	2.73	0.176	0.164	0.187
	DM or preDM	14	0.4	x	x	x	x
Deprivation score	1	658	18.4	3.08	0.148	0.142	0.154
	2	600	16.8	3.09	0.143	0.138	0.149
	3	790	22.1	3.05	0.159	0.153	0.164
	4	711	19.9	2.96	0.162	0.156	0.168
	5	435	12.2	2.95	0.148	0.141	0.155
	Missing	373	10.5	3.06	0.146	0.139	0.153
Region	East Midlands	55	1.5	3.16	0.104	0.089	0.119
	East of England	266	7.5	3.02	0.146	0.137	0.154
	London	222	6.2	2.92	0.164	0.153	0.175
	North East	71	2.0	2.53	0.146	0.127	0.165
	North West	366	10.3	2.86	0.160	0.152	0.168
	Northern Ireland	154	4.3	3.30	0.135	0.125	0.146
	Scotland	347	9.7	3.16	0.167	0.159	0.176
	South Central	331	9.3	3.17	0.127	0.120	0.134
	South East Coast	429	12.0	2.98	0.137	0.131	0.144
	South West	201	5.6	3.07	0.122	0.114	0.131
	Wales	450	12.6	3.14	0.155	0.148	0.161
	West Midlands	493	13.8	2.89	0.177	0.169	0.185
	Yorkshire & Humber	182	5.1	3.13	0.156	0.144	0.167
Alcohol consumption	Current drinker	2,200	61.7	3.07	0.152	0.149	0.155
	Ex drinker	61	1.7	3.04	0.160	0.142	0.179
	Teetotal	495	13.9	2.68	0.167	0.159	0.174

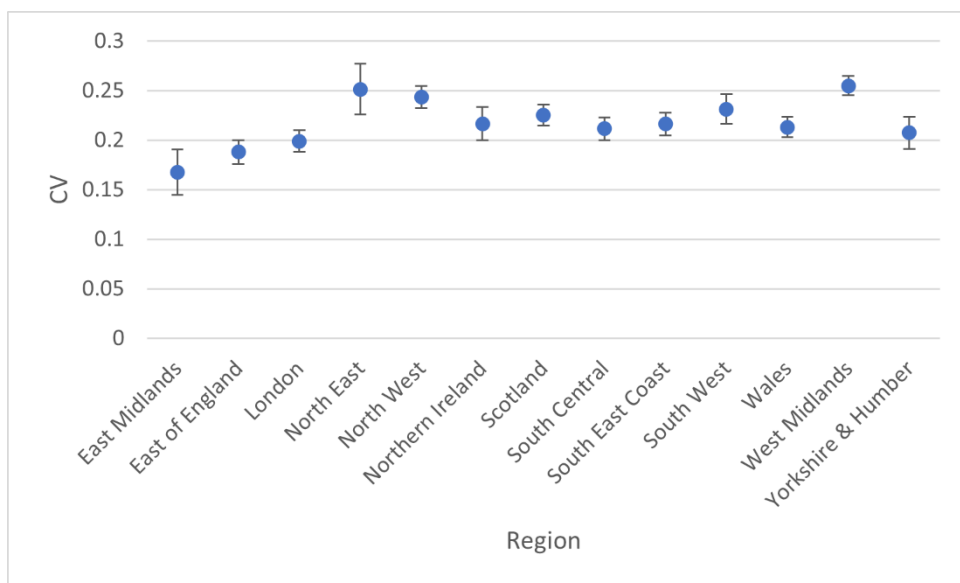


Figure 18.1 CVT by region for FEV1

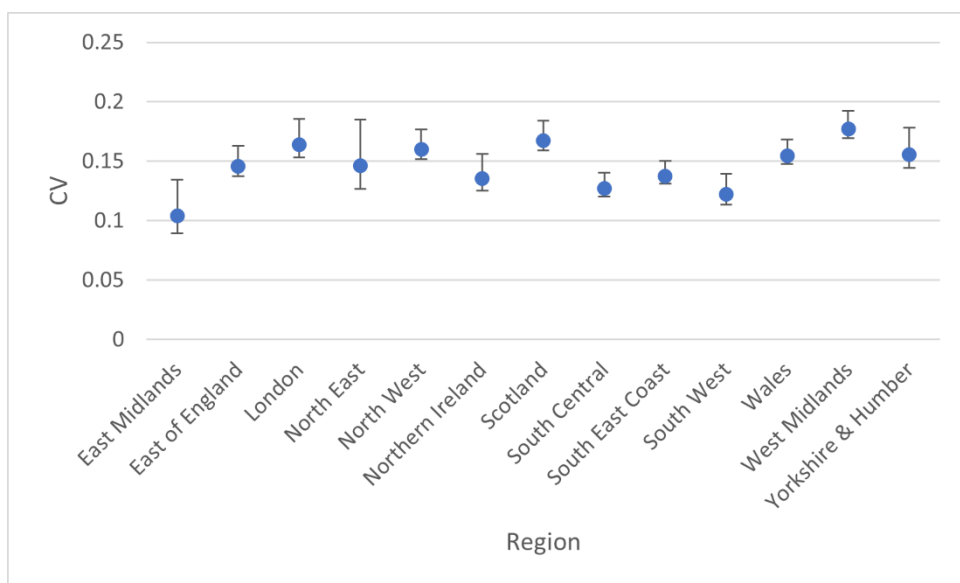


Figure 18.2 CVT by region for FEV1

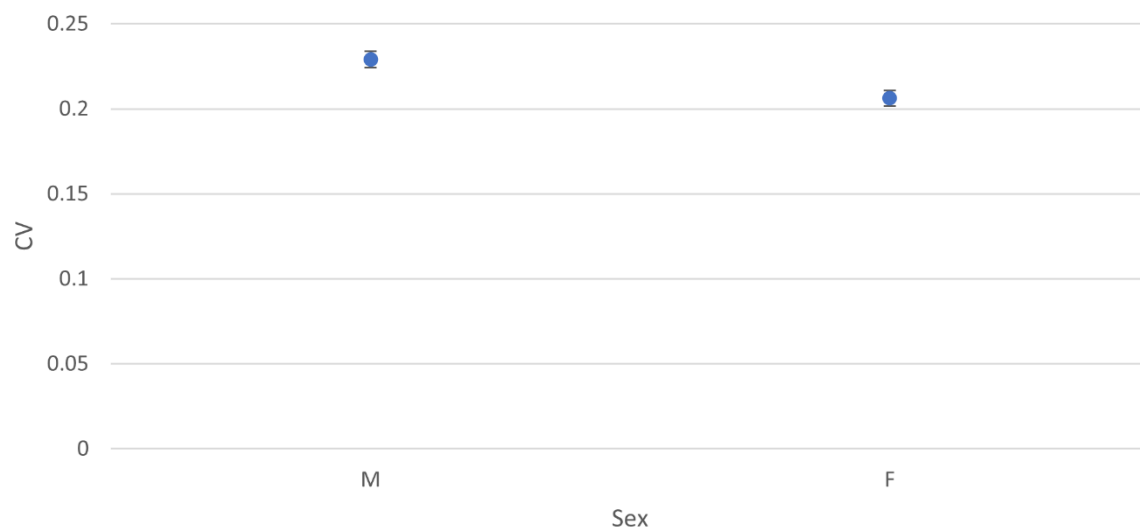


Figure 18.3 CVT by sex for FEV1

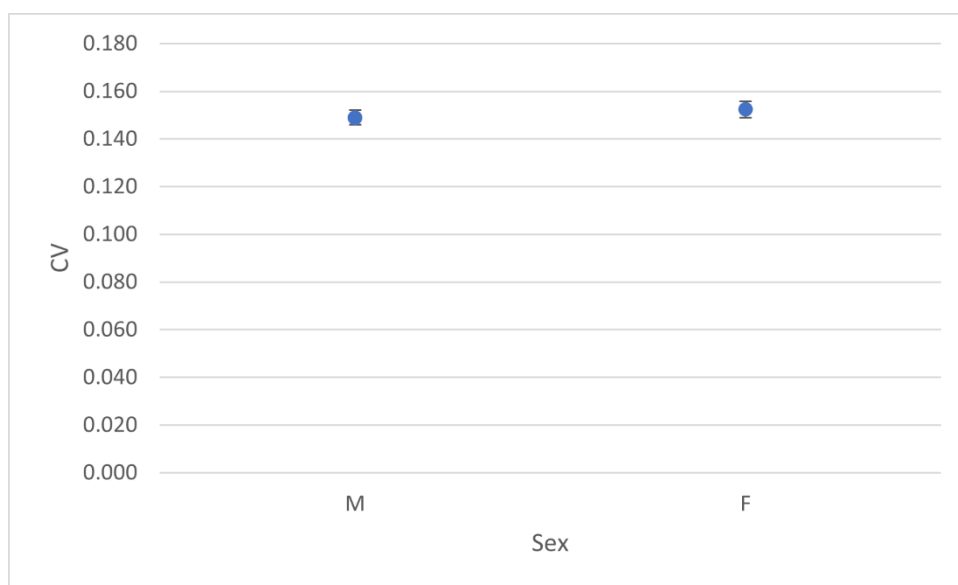


Figure 18.4 CVT by sex for FVC

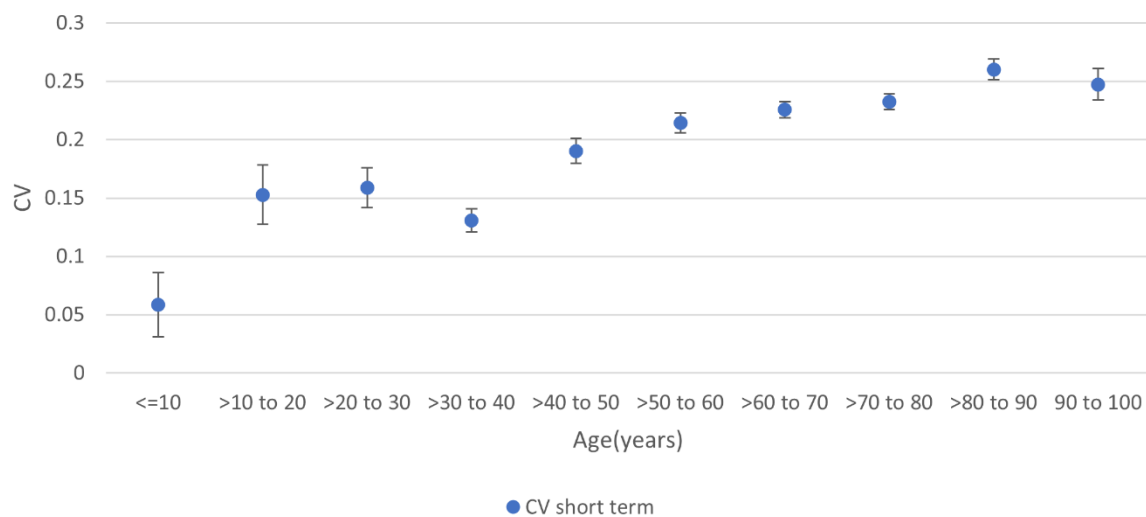


Figure 18.5 CVT by age for FEV1

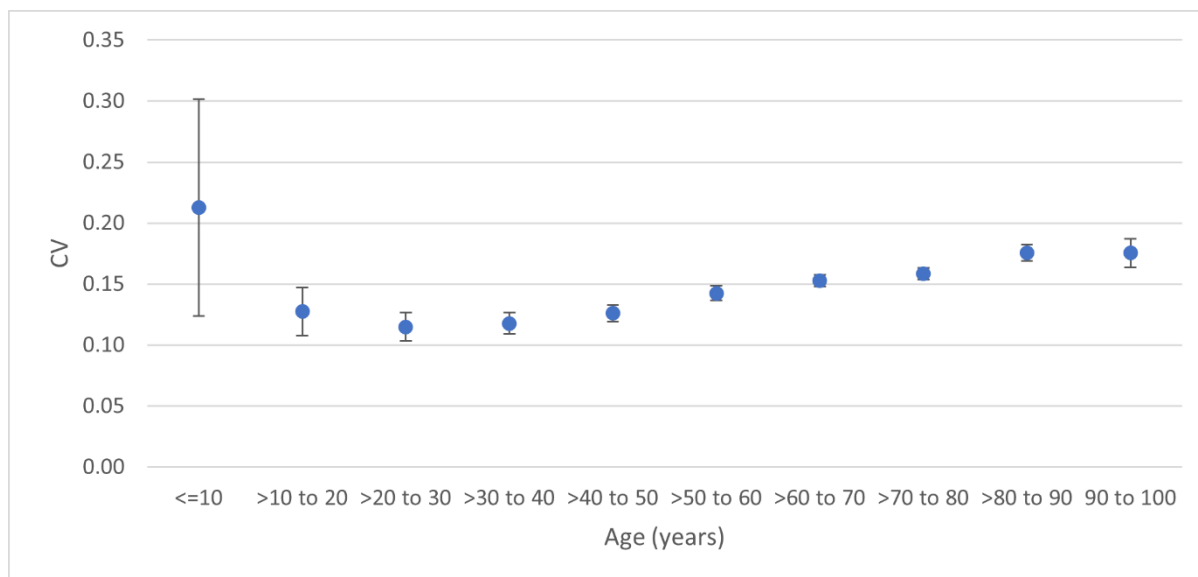


Figure 18.6 CVT by age for FVC

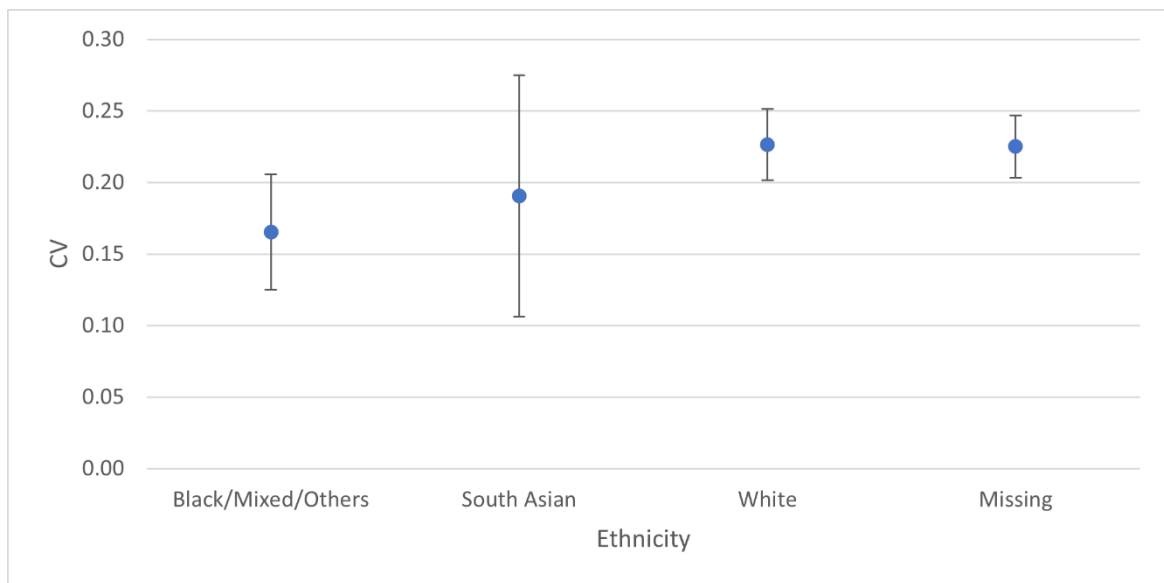


Figure 18.7 CVT by ethnicity for FEV1

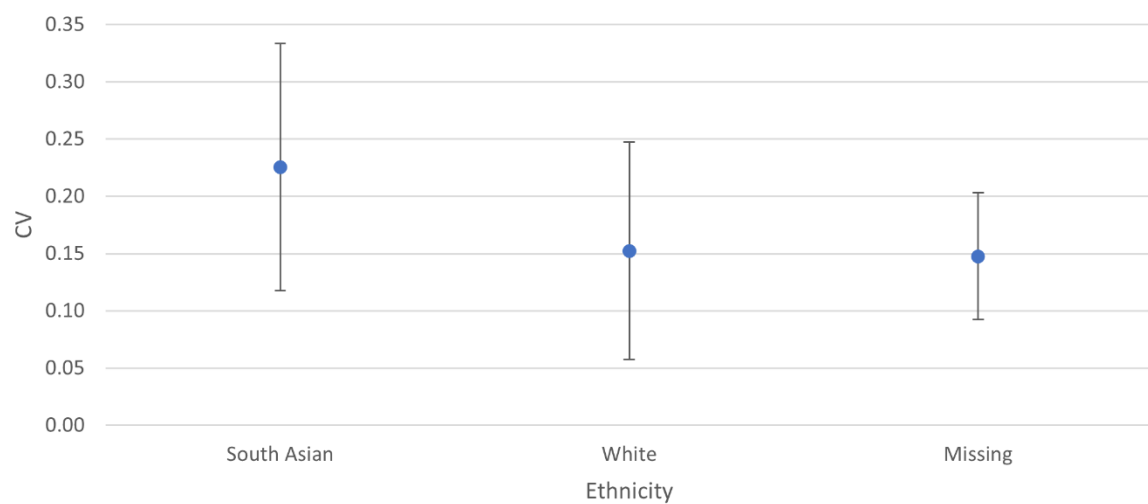


Figure 18.8 CVT by ethnicity for FVC

(Black/mixed/others ethnicity had < 50 participants so is omitted)

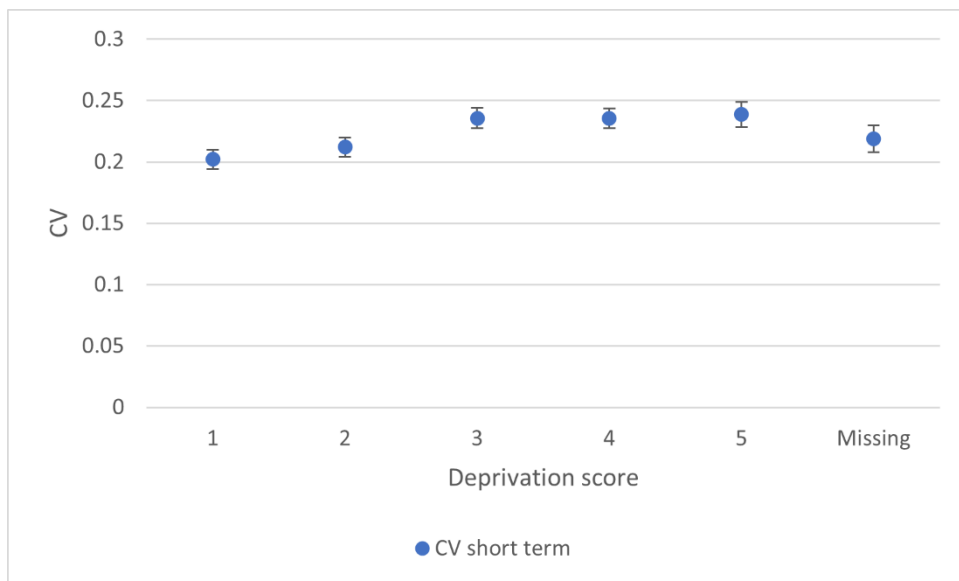


Figure 18.9 CVT by Townsend deprivation score for FEV1

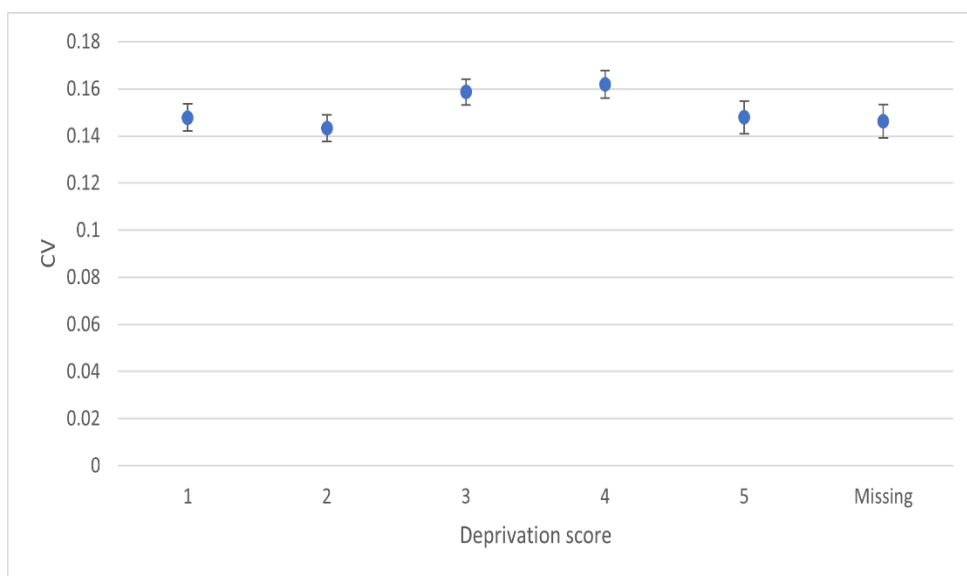


Figure 18.10 CVT by Townsend deprivation score for FVC

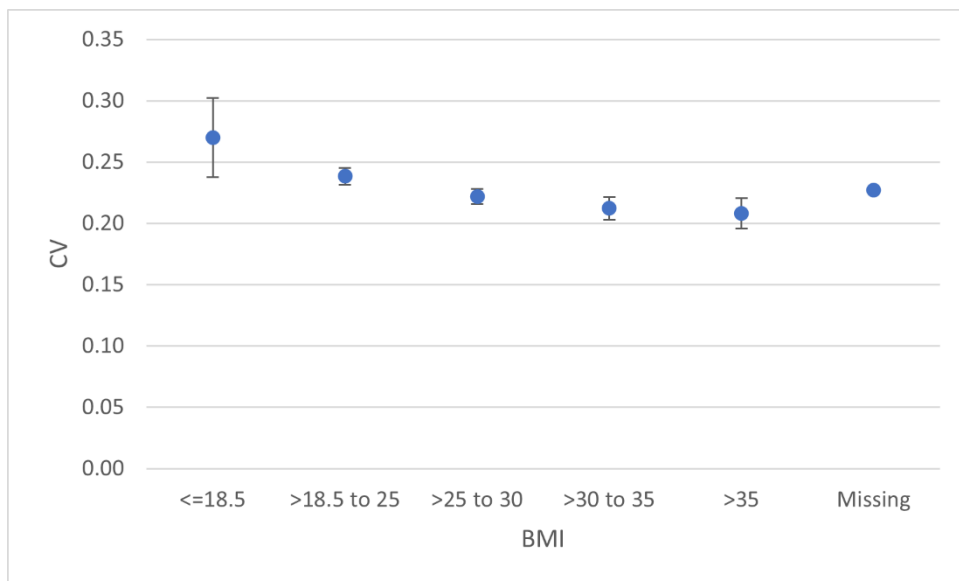


Figure 18.11 CVT by BMI for FEV1

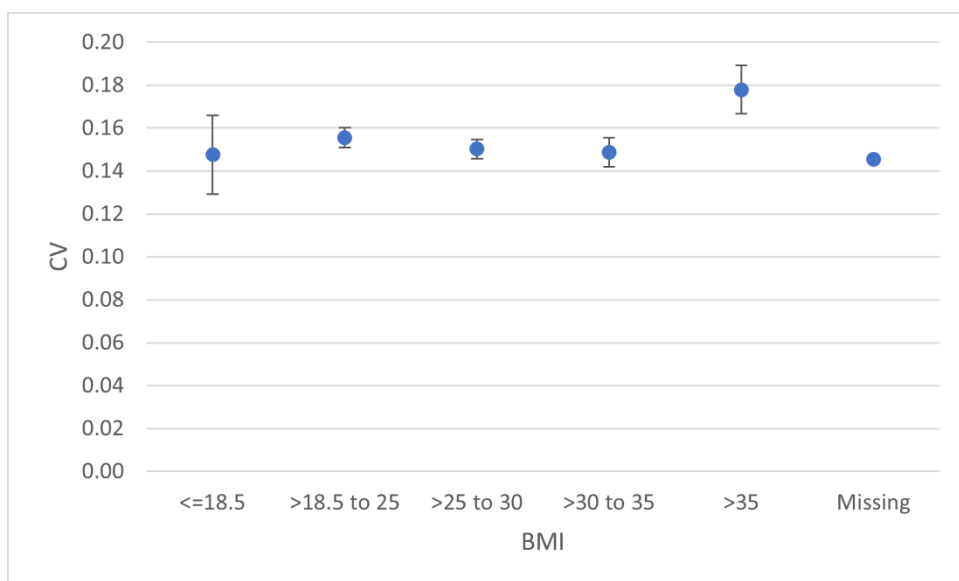


Figure 18.12 CVT by BMI for FVC

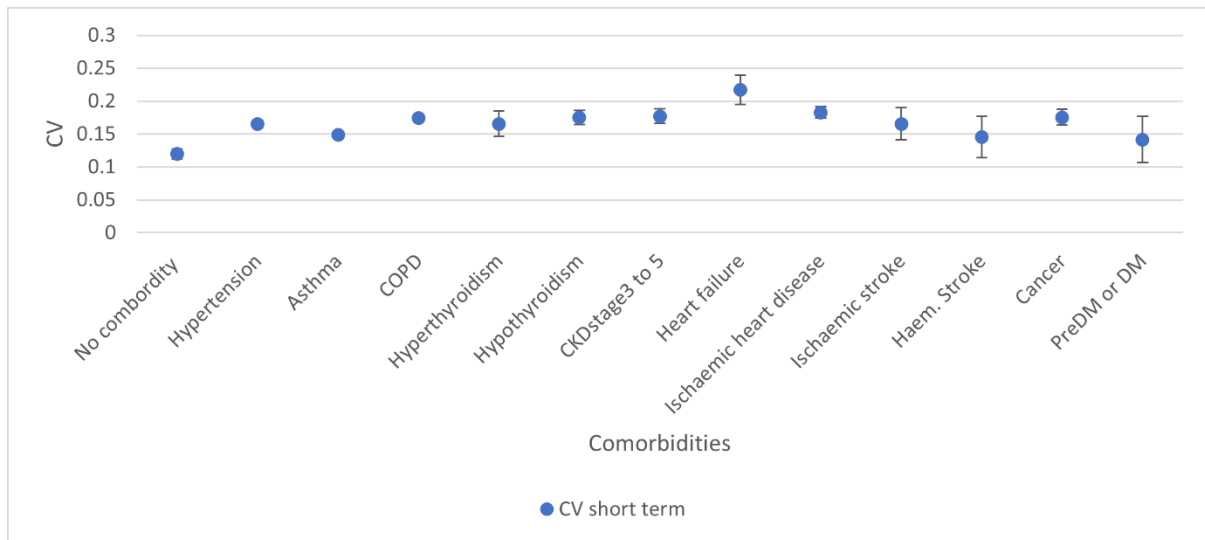


Figure 18.13 CVT by comorbidities for FEV1

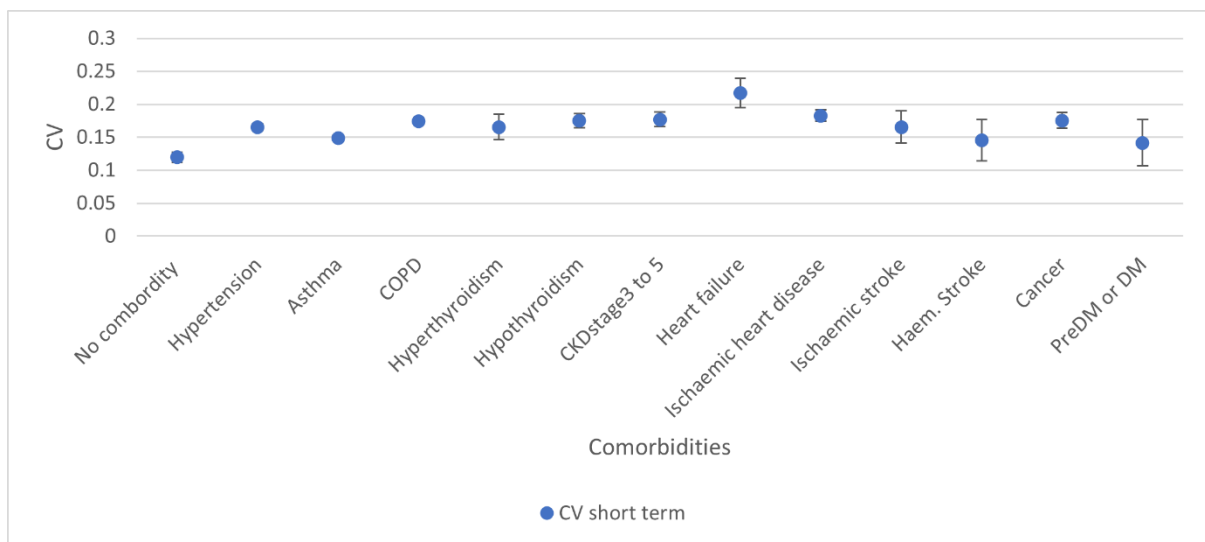


Figure 18.14 CVT by comorbidities for FVC



Figure 18.15 CVT by alcohol consumption for FEV1

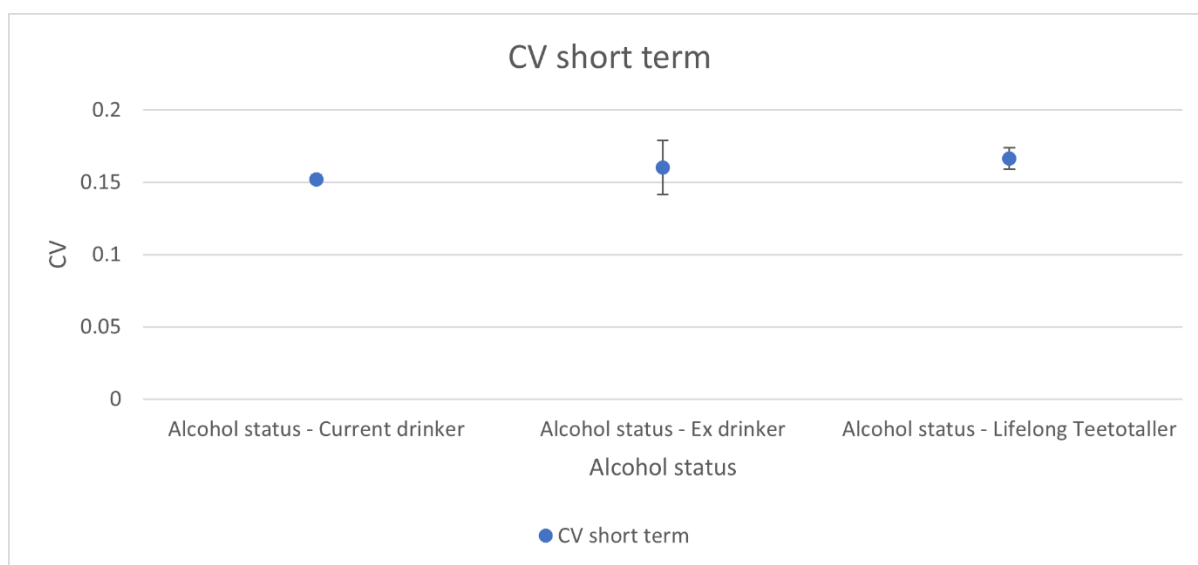


Figure 18.16 CVT by alcohol consumption for FVC

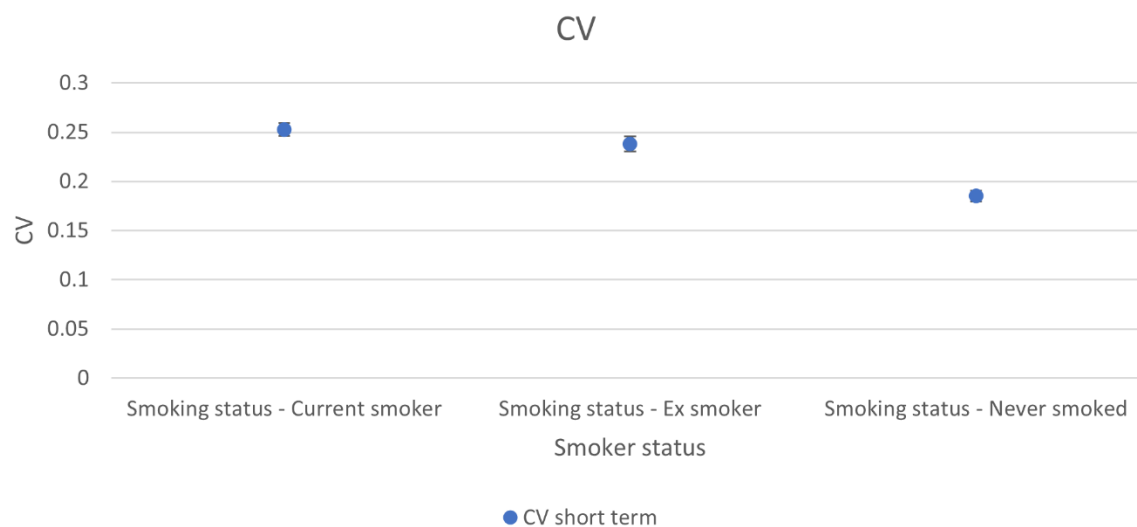


Figure 18.17 CVT by smoker status for FEV1

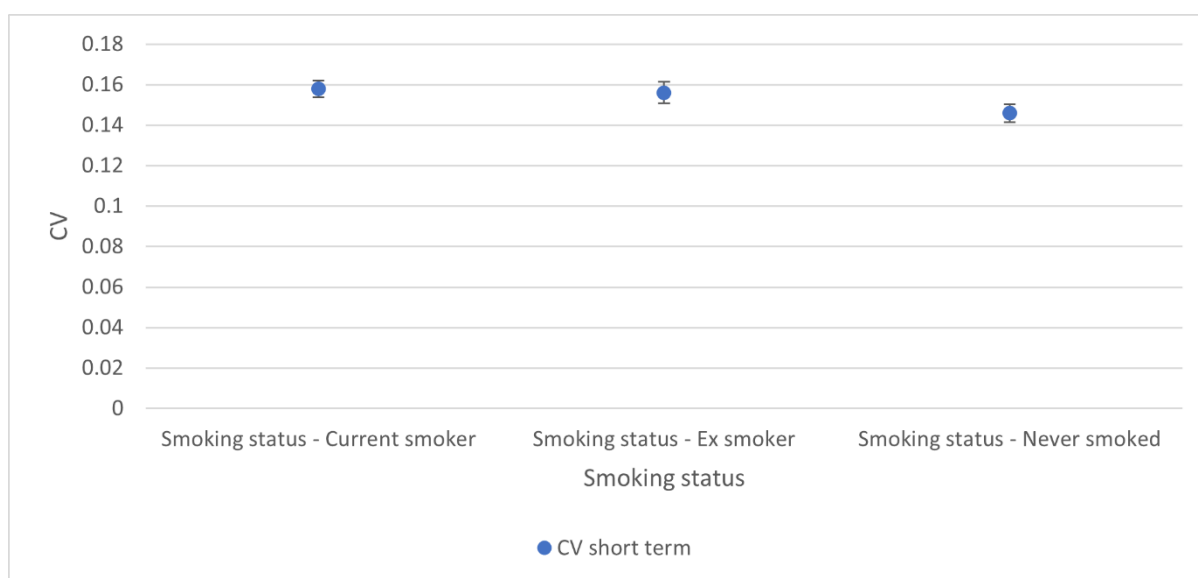


Figure 18.18 CVT by smoker status for FVC

18.2 Appendix 8.2 – sensitivity analyses

Table 18.3 Sensitivity analyses for FEV1

*=range

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
Year	2010	164	3.7	1.93	0.173	0.159	0.187
	2011	201	4.6	2.09	0.225	0.208	0.241
	2012	279	6.3	2.15	0.217	0.203	0.230
	2013	251	5.7	2.31	0.204	0.191	0.217
	2014	197	4.5	2.31	0.202	0.188	0.217
	2015	155	3.5	2.20	0.239	0.220	0.258
	2016	145	3.3	2.23	0.272	0.250	0.294
	2017	108	2.4	2.16	0.271	0.245	0.297
	2018	103	2.3	2.35	0.237	0.213	0.260
	2019	76	1.7	2.44	0.212	0.187	0.236
Quarter	1	376	8.5	2.12	0.247	0.235	0.260
	2	352	8.0	2.17	0.230	0.218	0.242
	3	314	7.1	2.14	0.237	0.223	0.251
	4	352	8.0	2.15	0.228	0.216	0.241
Number of measurements	2	57,572	N/A	1.97	0.191	0.190	0.192
	3	11,476	N/A	1.95	0.208	0.205	0.210
	4	3,284	N/A	2.03	0.222	0.218	0.226
	5	681	N/A	1.96	0.231	0.222	0.240
	6	262	N/A	2.10	0.226	0.213	0.239
	7	64	N/A	2.12	0.215	0.188	0.242
Patient median FEV1	0 to <0.5	9	0.2	x	x	x	x
	0.5 to < 1	348	7.9	0.88	0.387	0.370	0.405
	1 to <1.5	854	19.4	1.31	0.302	0.294	0.310
	1.5 to < 2	1,124	25.5	1.77	0.232	0.227	0.237
	2 to <2.5	962	21.8	2.22	0.210	0.205	0.214
	2.5 to < 3	575	13.0	2.68	0.186	0.181	0.191
	3 to <3.5	314	7.1	3.16	0.158	0.152	0.164
	3.5 to < 4	155	3.5	3.64	0.147	0.138	0.155
	4 to <4.5	58	1.3	4.08	0.153	0.140	0.167
Days between measurements	0 to < 4	1,503	34.1	2.26	0.256	0.250	0.262
	4 to < 7	88	2.0	2.22	0.235	0.208	0.261
	7 to < 14	399	9.0	1.97	0.230	0.219	0.242
	14 to < 21	498	11.3	1.94	0.206	0.196	0.215
	21 to < 28	458	10.4	1.89	0.174	0.165	0.183
	28 to < 60	1,341	30.4	1.87	0.179	0.174	0.185
	60 to < 90	125	2.8	1.78	0.218	0.196	0.240
Method	Linear regression	4412	100	2.03	0.224	0.221	0.228
	Arithmetic	4412	100	2.03	0.181	0.002*	1.585*
Unit	L	4412	100	2.03	0.224	0.221	0.228
	% predicted	515	N/A	66.97	0.160	0.154	0.167

Table 18.4 Sensitivity analyses for FVC.

*=range

		N	%	Mean	CV	95% LCI	95% UCI
Year	2010	101	2.8	2.95	0.166	0.150	0.182
	2011	191	5.4	3.06	0.138	0.128	0.148
	2012	230	6.4	3.09	0.144	0.136	0.153
	2013	211	5.9	3.21	0.119	0.112	0.127
	2014	185	5.2	3.25	0.138	0.128	0.148
	2015	147	4.1	3.21	0.144	0.133	0.155
	2016	161	4.5	3.30	0.183	0.170	0.197
	2017	141	4.0	3.22	0.150	0.138	0.161
	2018	154	4.3	3.23	0.154	0.143	0.166
	2019	149	4.2	3.33	0.158	0.146	0.170
Quarter	1	359	10.1	3.08	0.165	0.157	0.173
	2	331	9.3	3.11	0.156	0.148	0.164
	3	332	9.3	3.07	0.138	0.131	0.146
	4	319	8.9	3.17	0.154	0.146	0.163
Number of measurements	2	40,301	N/A	2.90	0.138	0.137	0.139
	3	7,772	N/A	2.91	0.146	0.144	0.148
	4	2,488	N/A	3.00	0.152	0.149	0.156
	5	622	N/A	3.08	0.150	0.145	0.156
	6	278	N/A	3.15	0.152	0.145	0.160
	7	68	N/A	3.25	0.155	0.139	0.171
Patient median FVC	0 to <0.5	0	0	x	x	x	x
	0.5 to < 1	22	0.6	x	x	x	x
	1 to <1.5	123	3.4	1.36	0.262	0.243	0.280
	1.5 to < 2	366	10.3	1.82	0.208	0.199	0.216
	2 to <2.5	656	18.4	2.26	0.175	0.170	0.180
	2.5 to < 3	680	19.1	2.74	0.175	0.171	0.180
	3 to <3.5	621	17.4	3.22	0.151	0.147	0.155
	3.5 to < 4	490	13.7	3.70	0.131	0.127	0.135

		N	%	Mean	CV	95% LCI	95% UCI
	4 to <4.5	315	8.8	4.20	0.107	0.103	0.112
	4.5 to < 5	158	4.4	4.65	0.102	0.096	0.107
	5 to < 5.5	89	2.5	5.18	0.094	0.087	0.101
Days between measurements	0 to < 4	1,784	50.0	3.18	0.161	0.158	0.165
	4 to < 7	63	1.8	2.85	0.162	0.144	0.181
	7 to < 14	265	7.4	2.97	0.150	0.141	0.158
	14 to < 21	297	8.3	2.85	0.145	0.137	0.153
	21 to < 28	279	7.8	2.86	0.133	0.125	0.141
	28 to < 60	808	22.7	2.88	0.133	0.128	0.137
	60 to < 90	71	2.0	2.77	0.125	0.109	0.140
Method	Linear regression	3,567	100	3.03	0.152	0.150	0.155
	Arithmetic	3567	100	3.04	0.131	0*	0.805*
Unit	L	3,567	100	3.03	0.152	0.150	0.155
	% predicted	100	N/A	81.03	0.128	0.116	0.140

Table 18.5 Results for long-term data (i.e. all measurements included) for FEV1

		N	%	Mean	CV	95% LCI	95% UCI
All		129,161	100	1.72	0.228	0.227	0.229
Sex	Male	67,332	52.1	1.96	0.228	0.227	0.229
	Female	61,829	47.9	1.46	0.219	0.218	0.220
Age	>10 to 20	384	0.3	2.21	0.234	0.224	0.244
	>20 to 30	861	0.7	2.77	0.190	0.185	0.196
	>30 to 40	1,120	0.9	2.96	0.150	0.147	0.154
	>40 to 50	3,169	2.5	2.66	0.179	0.176	0.181
	>50 to 60	10,828	8.4	2.26	0.202	0.201	0.204
	>60 to 70	24,690	19.1	1.91	0.225	0.224	0.226
	>70 to 80	41,666	32.3	1.68	0.233	0.232	0.234
	>80 to 90	33,762	26.1	1.46	0.239	0.238	0.240
	90 to 100	12,049	9.3	1.27	0.252	0.250	0.254
Ethnicity	Black	448	0.3	1.79	0.233	0.223	0.244
	Mixed	161	0.1	1.84	0.255	0.236	0.274
	Others	482	0.4	1.80	0.220	0.210	0.229
	South Asian	1,054	0.8	1.69	0.241	0.234	0.248
	White	69,563	53.9	1.75	0.225	0.224	0.226
	Missing	57,453	44.5	1.68	0.231	0.230	0.232
BMI	<18.5	3,100	2.4	1.46	0.279	0.273	0.284
	18.5 to <25	41,225	31.9	1.66	0.241	0.240	0.242
	25 to <30	37,338	28.9	1.76	0.218	0.217	0.219
	30 to <35	15,867	12.3	1.77	0.216	0.214	0.217
	>35	7,689	6.0	1.71	0.210	0.207	0.212
	Missing	23,942	18.5	1.77	0.230	0.228	0.231
Smoker status	Current smoker	54,790	42.4	1.69	0.242	0.241	0.243
	Ex-smoker	36,448	28.2	1.65	0.227	0.225	0.228
	Never smoked	24,364	18.9	1.83	0.200	0.199	0.202
	Missing	13,559	10.5				
Respiratory status	Asthma	68,964	53.39383	1.71	0.229	0.228	0.230
	COPD	99,609	77.12003	1.58	0.249	0.248	0.250

Table 18.6 Results for long-term data (i.e. all measurements included) for FVC

		N	%	Mean	CV	95% LCI	95% UCI
All		75,956	100	2.73	0.164	0.164	0.165
Sex	Male	39,722	52.3	3.18	0.162	0.161	0.163
	Female	36,234	47.7	2.23	0.160	0.160	0.161
Age	>10 to 20	338	0.4	2.63	0.237	0.226	0.248
	>20 to 30	725	1.0	3.44	0.180	0.175	0.186
	>30 to 40	935	1.2	3.86	0.119	0.116	0.123
	>40 to 50	2,244	3.0	3.64	0.132	0.130	0.135
	>50 to 60	6,787	8.9	3.30	0.144	0.142	0.145
	>60 to 70	14,912	19.6	2.97	0.158	0.157	0.159
	>70 to 80	24,546	32.3	2.70	0.164	0.163	0.165
	>80 to 90	19,013	25.0	2.40	0.176	0.175	0.177
	90 to 100	6,154	8.1	2.13	0.191	0.189	0.193
	Missing						
Ethnicity	Black	308	0.4	2.57	0.210	0.200	0.221
	Mixed	94	0.1	2.77	0.181	0.163	0.198
	Others	288	0.4	2.64	0.175	0.166	0.185
	South Asian	676	0.9	2.41	0.204	0.197	0.211
	White	40,702	53.6	2.76	0.163	0.162	0.164
	Missing	33,888	44.6	2.70	0.164	0.163	0.165
BMI	<18.5	1,704	2.2	2.51	0.177	0.173	0.181
	18.5 to <25	24,096	31.7	2.73	0.165	0.164	0.166
	25 to <30	22,061	29.0	2.75	0.162	0.161	0.163
	30 to <35	9,216	12.1	2.66	0.162	0.160	0.163
	>35	4,460	5.9	2.51	0.168	0.166	0.170
	Missing	14,419	19.0	2.84	0.165	0.164	0.166
Smoker status	Current smoker	31,188	41.1	2.75	0.166	0.166	0.167
	Ex-smoker	20,625	27.2	2.67	0.165	0.164	0.166
	Never smoked	15,553	20.5	2.71	0.157	0.155	0.158
	Missing	8590	11.3				
Respiratory status	Asthma	41,639	54.8	2.70	0.166	0.165	0.167
	COPD	53,824	70.9	2.62	0.171	0.171	0.172

	Respiratory disease status	Median time between measurements (days)	BMI				
Sex				M	M	F	F
Age				20-60	>60	20-60	>60
	No respiratory disease	<180	<30	0.132	0.148	0.127	0.134
			>30	0.150	0.154	0.134	0.157
		>180	<30	0.096	0.135	0.183	0.130
			>30	0.133	0.129	0.138	0.133
	Asthma	<180	<30	0.204	0.275	0.193	0.249
			>30	0.203	0.242	0.189	0.251
		>180	<30	0.188	0.236	0.175	0.230
			>30	0.189	0.215	0.181	0.207
	COPD	<180	<30	0.284	0.312	0.245	0.293
			>30	0.247	0.266	0.269	0.227
		>180	<30	0.229	0.241	0.215	0.242
			>30	0.213	0.223	0.210	0.218

Figure 18.19 Heat map including time between measurements for long-term FEV1

	Respiratory disease status	Median time between measurements (days)	BMI				
Sex				M	M	F	F
Age				20-60	>60	20-60	>60
	No respiratory disease	<180	<30	0.106	0.127	0.135	0.138
			>30	0.125	0.147	0.113	0.193
		>180	<30	0.096	0.133	0.145	0.132
			>30	0.128	0.123	0.146	0.135
	Asthma	<180	<30	0.131	0.184	0.141	0.172
			>30	0.157	0.181	0.152	0.172
		>180	<30	0.131	0.166	0.141	0.166
			>30	0.134	0.162	0.151	0.161
	COPD	<180	<30	0.156	0.185	0.153	0.174
			>30	0.170	0.182	0.177	0.166
		>180	<30	0.144	0.166	0.154	0.169
			>30	0.150	0.163	0.165	0.162

Figure 18.20 Heat map including time between measurements for long-term FVC

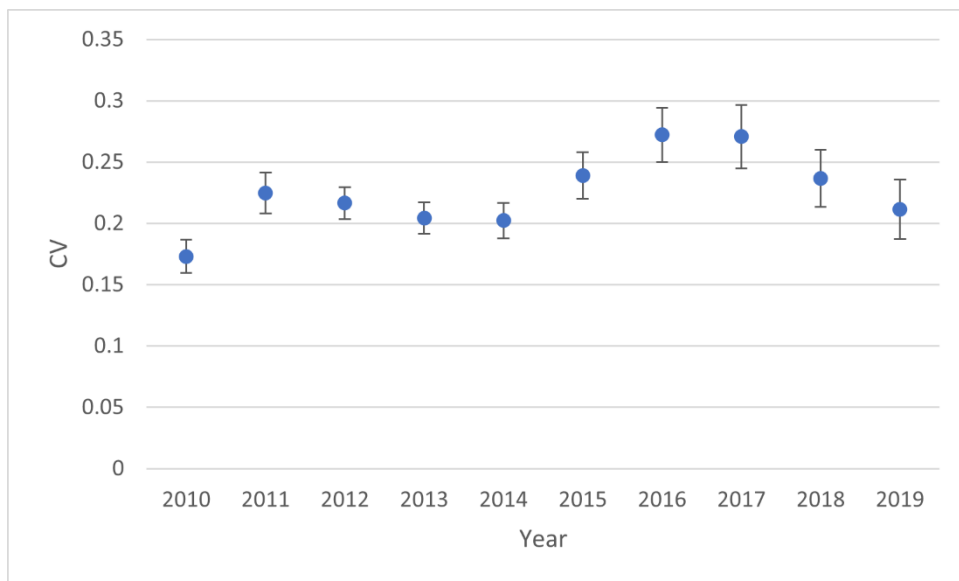


Figure 18.21 CVT by year for FEV1

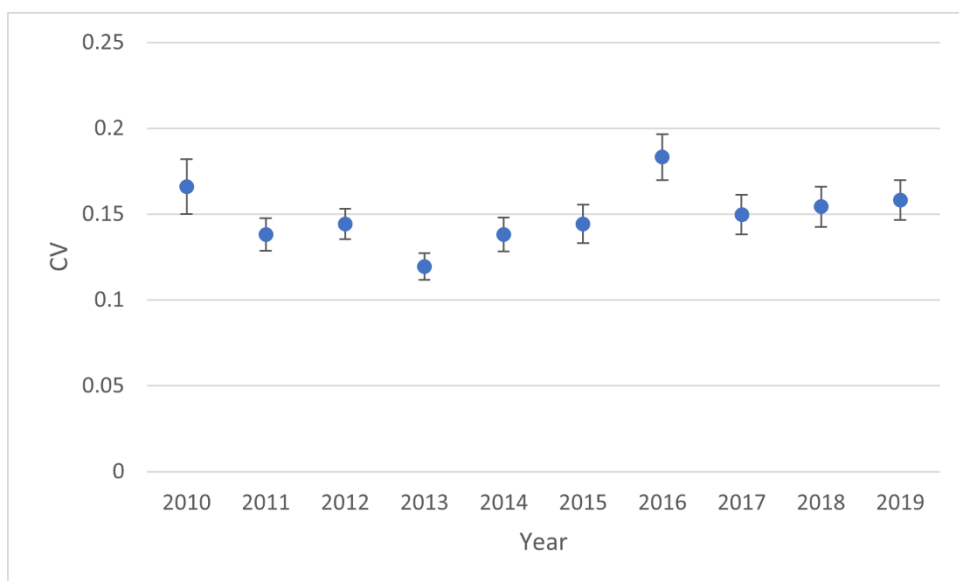


Figure 18.22 CVT by year for FVC

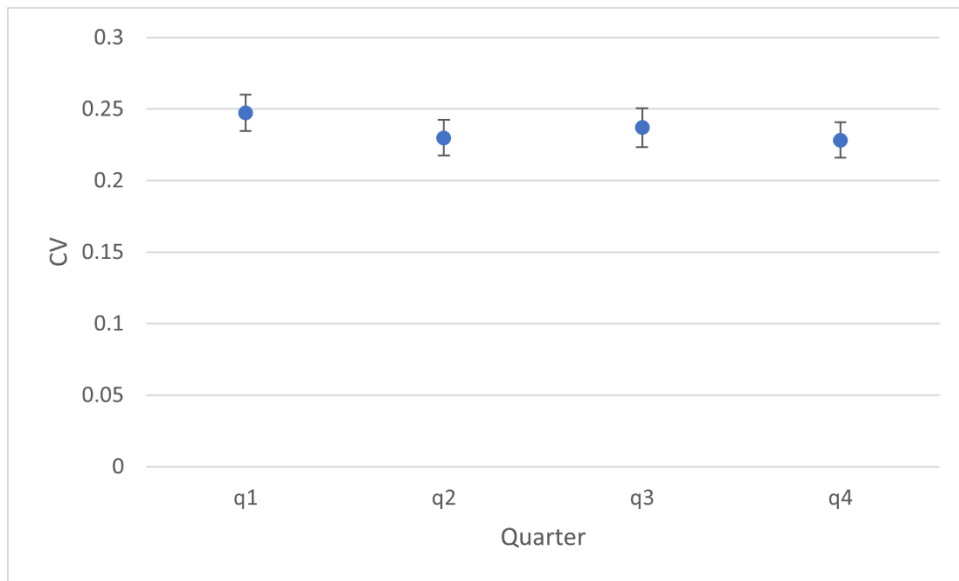


Figure 18.23 CVT by quarter for FEV1

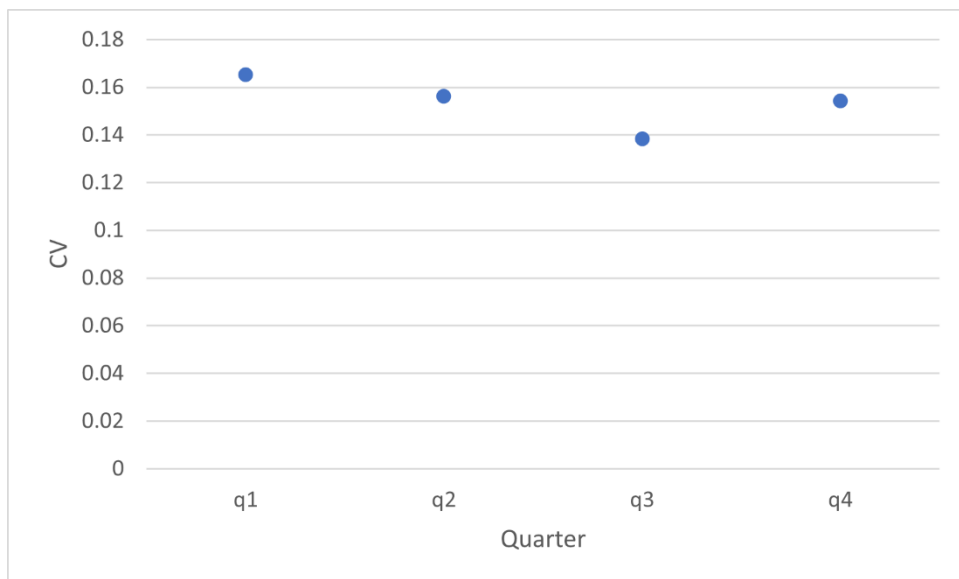


Figure 18.24 CVT by quarter for FVC

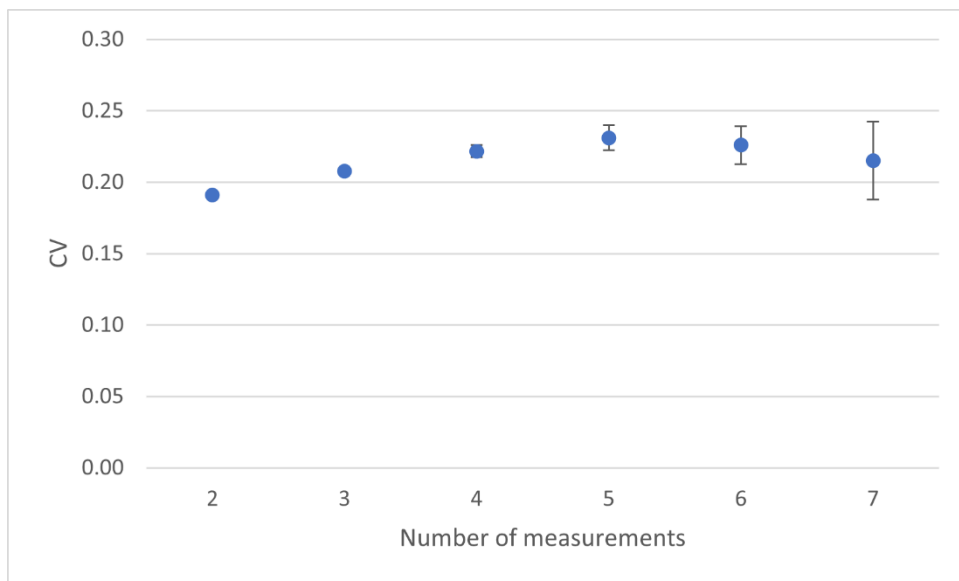


Figure 18.25 CVT by Number of measurements for FEV1

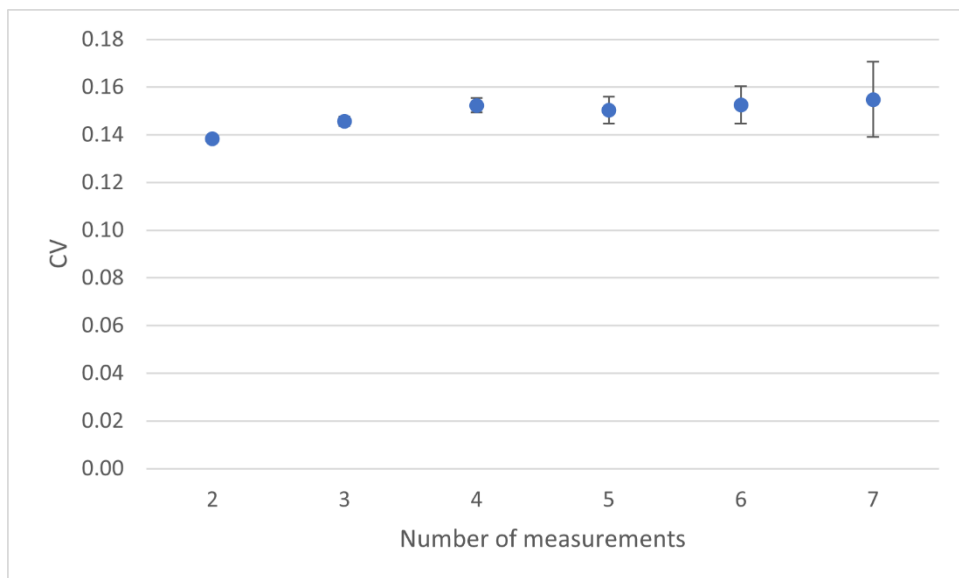


Figure 18.26 CVT by Number of measurements for FVC

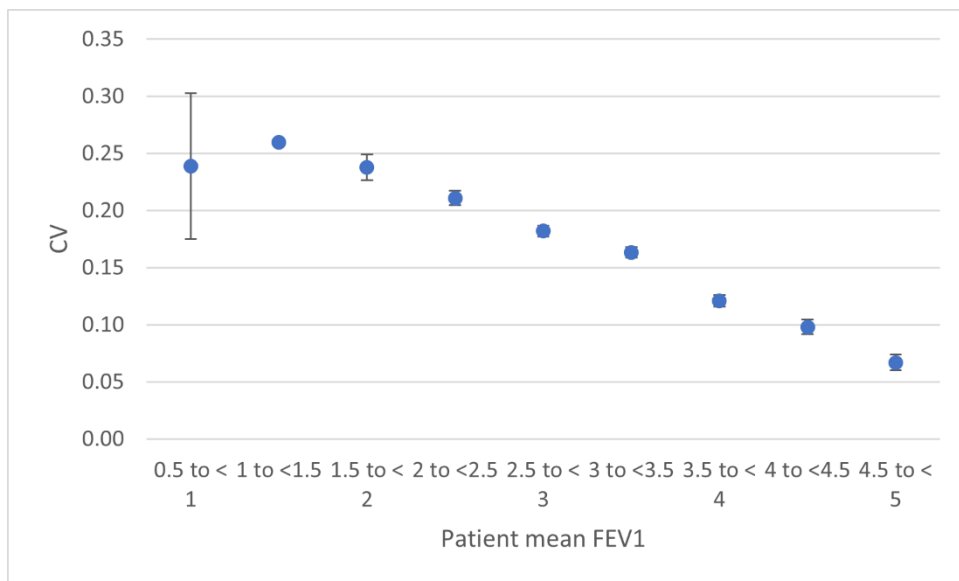


Figure 18.27 CVT by mean for FEV1

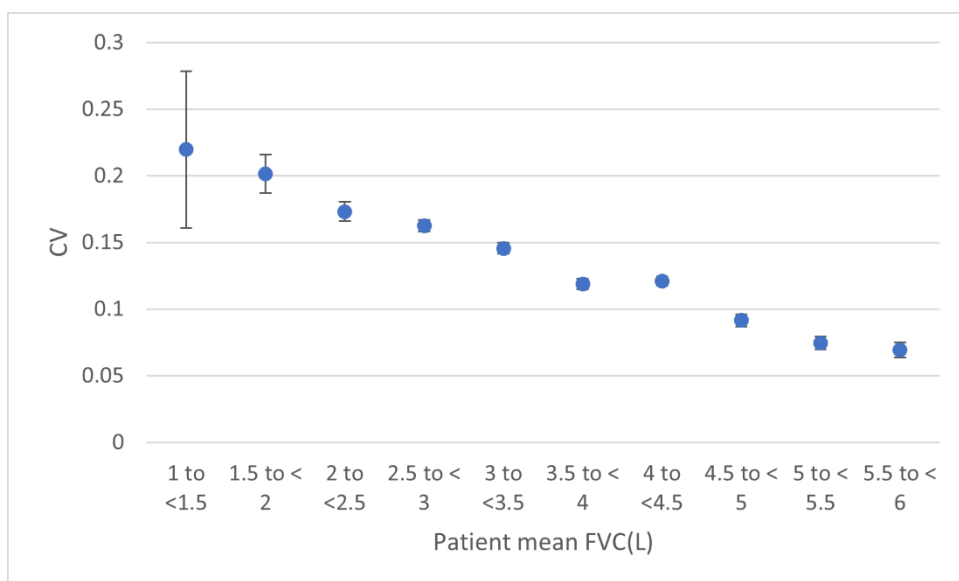


Figure 18.28 CVT by Mean for FVC

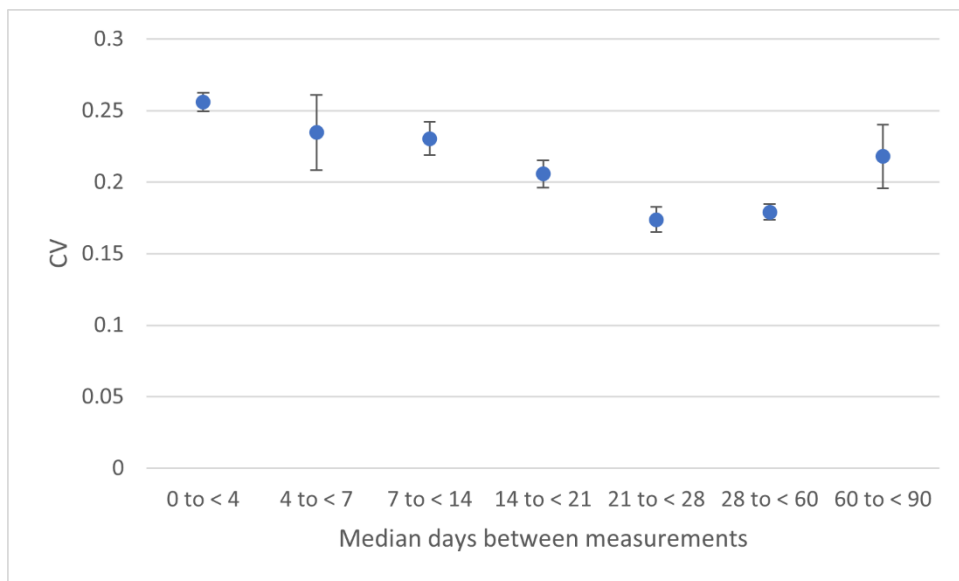


Figure 18.29 CVT by days between measurements for FEV1

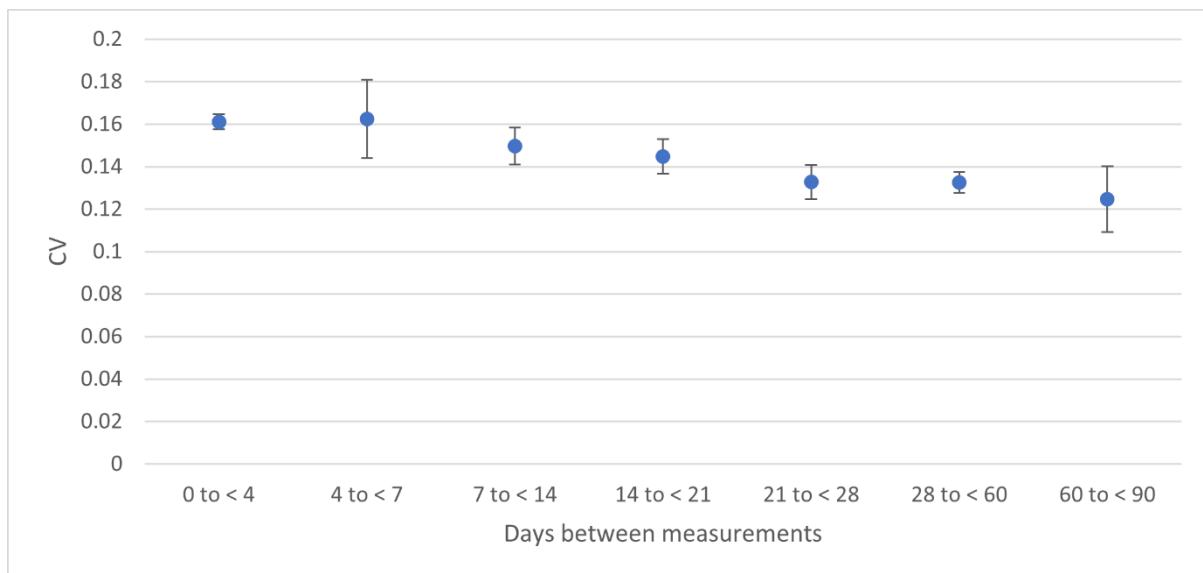


Figure 18.30 CVT by days between measurements for FVC

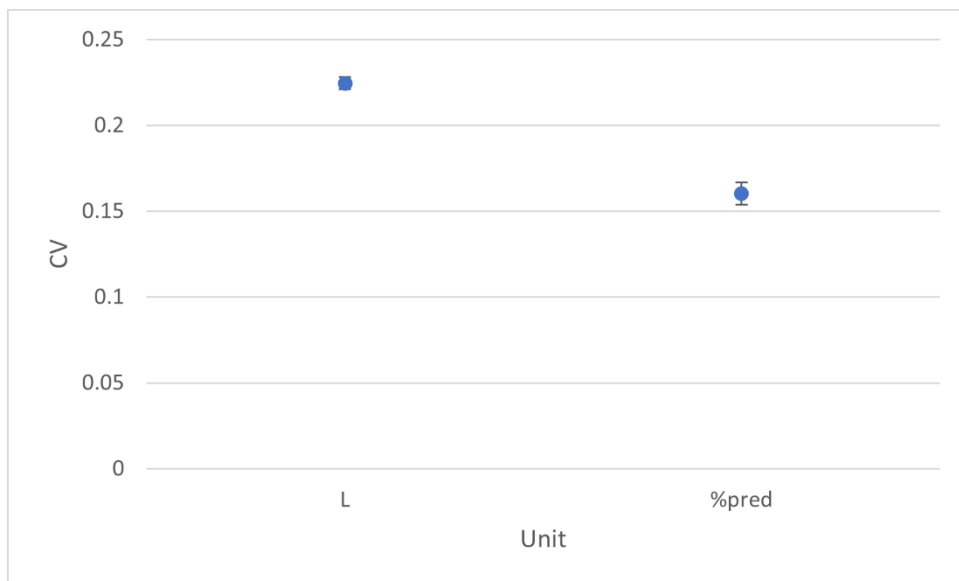


Figure 18.31 CVT by unit of measurement for FEV1

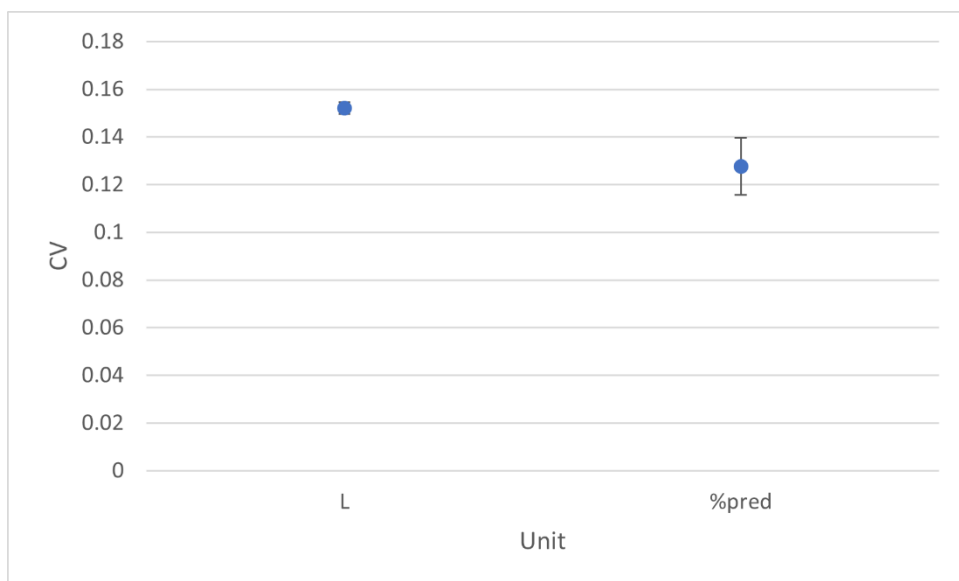


Figure 18.32 CVT by unit of measurement for FVC

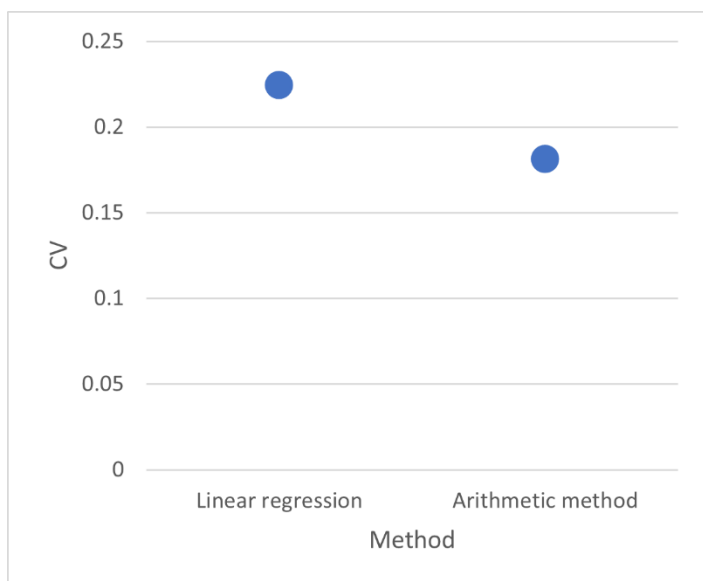


Figure 18.33 CVT by arithmetic calculation methods for FEV1

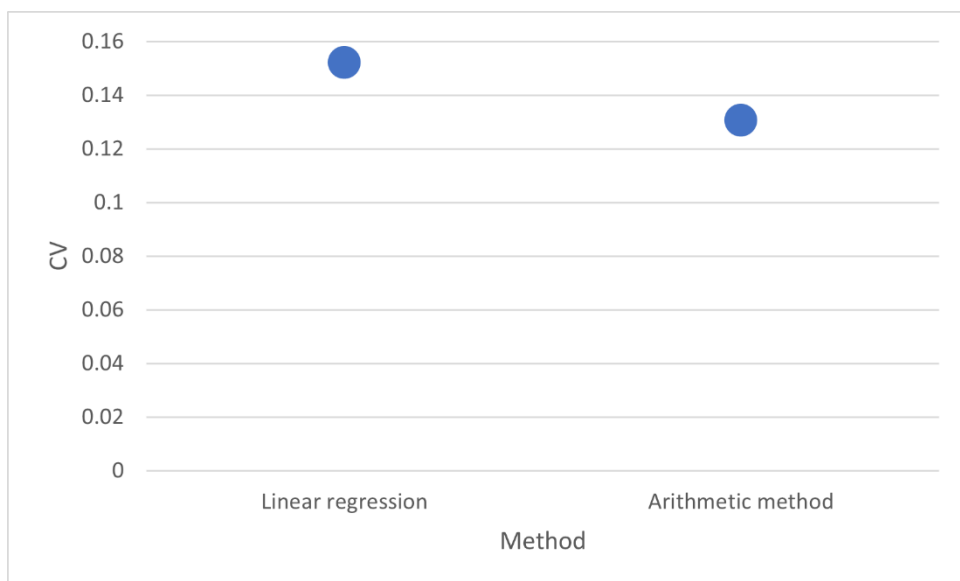


Figure 18.34 CVT by arithmetic calculation methods for FVC

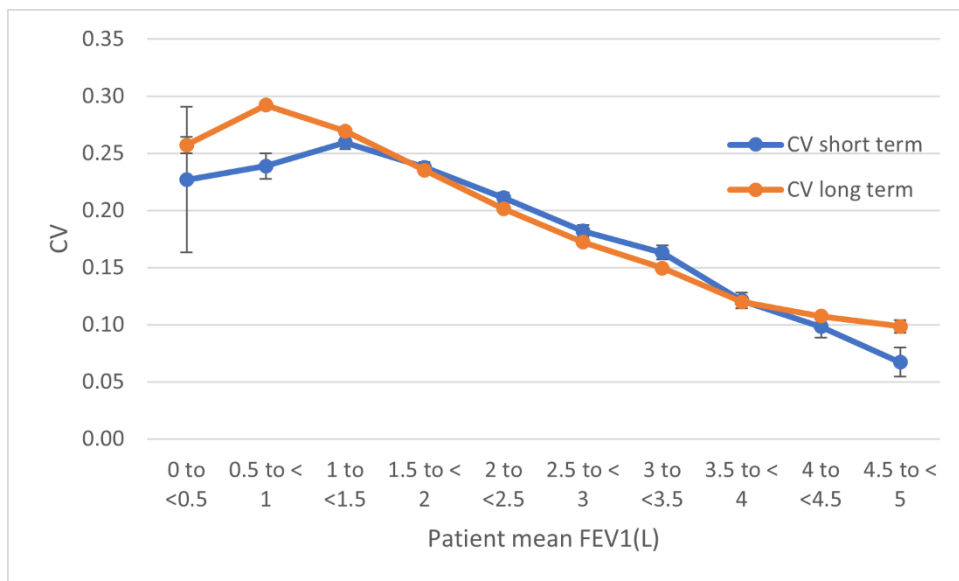


Figure 18.35 CVT by mean short and long-term data for FEV1

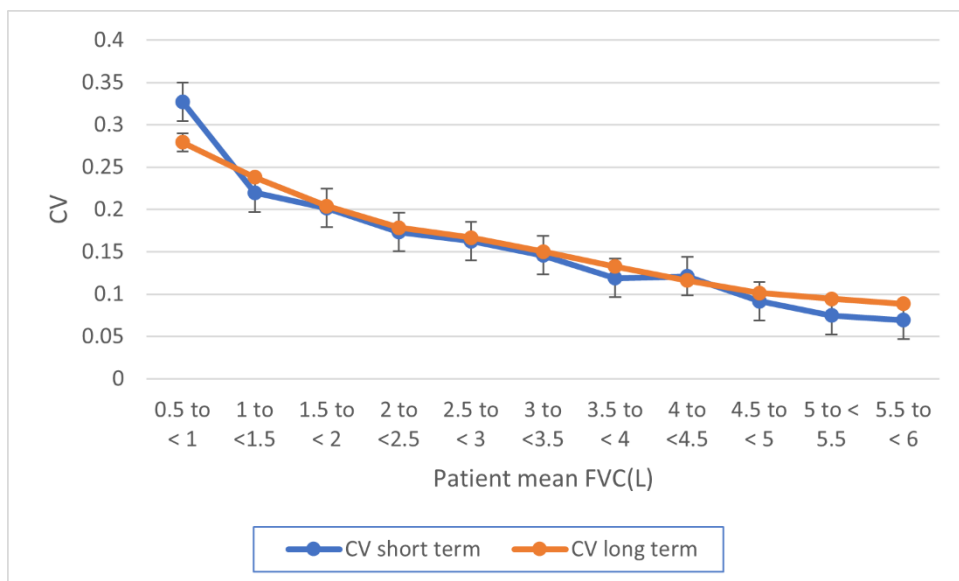


Figure 18.36 CVT by mean short and long-term data for FVC

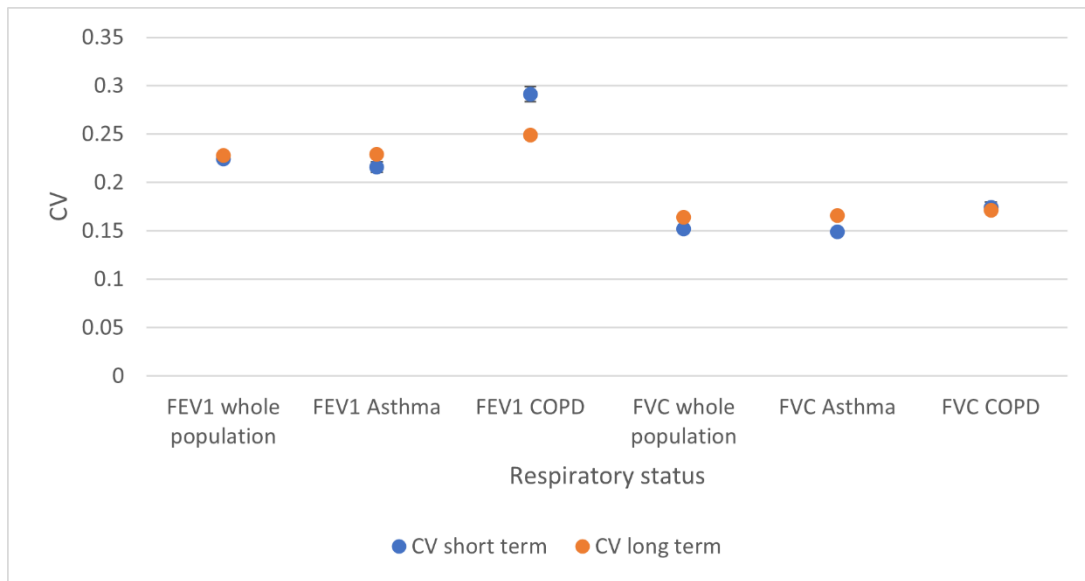


Figure 18.37 CVT by respiratory status for short and long-term data for FEV1 and FVC

19 Appendix 9 Depression measurement instrument

appendices

19.1 Appendix 9.1 - Subgroup analysis

Table 19.1 Other subgroup analyses.

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
Region	East Midlands	6,049	3.1	13.54	0.396	0.391	0.401
	East of England	8,579	4.4	9.85	0.512	0.503	0.520
	London	38,533	19.7	11.12	0.456	0.452	0.459
	North East	8,004	4.1	13.50	0.393	0.388	0.398
	North West	42,231	21.6	13.14	0.389	0.388	0.391
	Northern Ireland	579	0.3	13.74	0.387	0.372	0.402
	South East	35,365	18.1	13.34	0.408	0.406	0.410
	South West	15,087	7.7	13.71	0.388	0.385	0.392
	West Midlands	30,862	15.8	12.95	0.413	0.411	0.416
	Yorkshire & Humber	9,792	5.0	13.44	0.383	0.379	0.387
	Missing	90	0.0	-	-	-	-
Alcohol consumption (units/week)	0	41,555	21.3	12.74	0.408	0.406	0.411
	>0 to 10	53,447	27.4	12.62	0.421	0.419	0.423
	>10 to 20	11,628	6.0	12.55	0.422	0.417	0.426
	>20 to 30	4,015	2.1	12.74	0.406	0.399	0.413
	>30 to 40	1,265	0.6	13.22	0.399	0.387	0.411
	>40 to 50	715	0.4	13.55	0.388	0.372	0.403
	>50 to 60	483	0.2	14.08	0.377	0.360	0.394
	>60 to 70	302	0.2	15.10	0.352	0.333	0.371
	Missing	81,761	41.9	12.74	0.408	0.406	0.411
	Missing	81,761	41.9	12.74	0.408	0.406	0.411
Smoker Status	Current Smoker	86,735	43.6*	13.45	0.395	0.393	0.396
	Ex-smoker	33,811	17.0*	11.94	0.439	0.437	0.442
	Never smoked	78,437	39.4*	12.36	0.424	0.423	0.426
	Uncertain	35,714	18.3*	12.05	0.436	0.433	0.439

* Errors exist in the data for categorisation of smoking status since the numbers in the groups sum to more than 100%

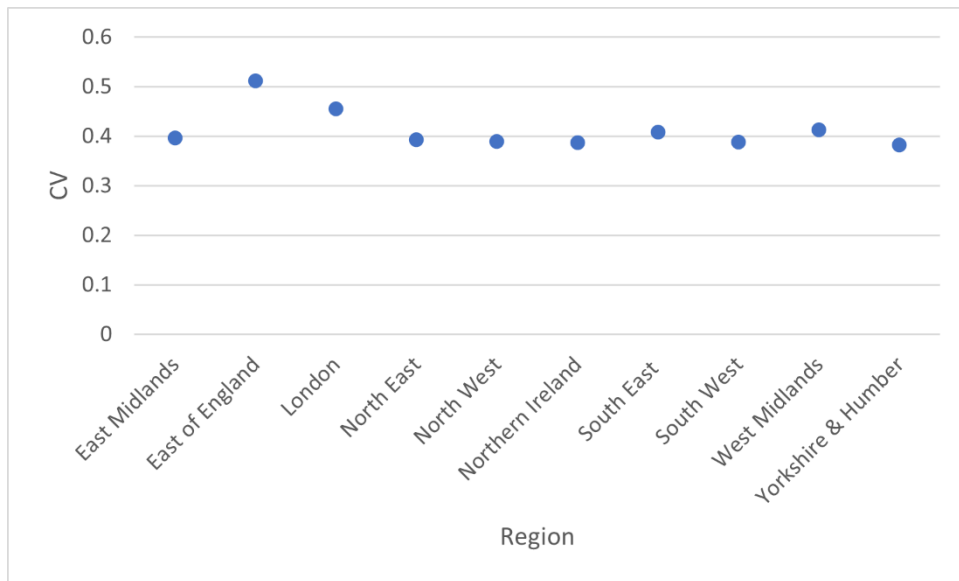


Figure 19.1 CVT by region for PHQ-9

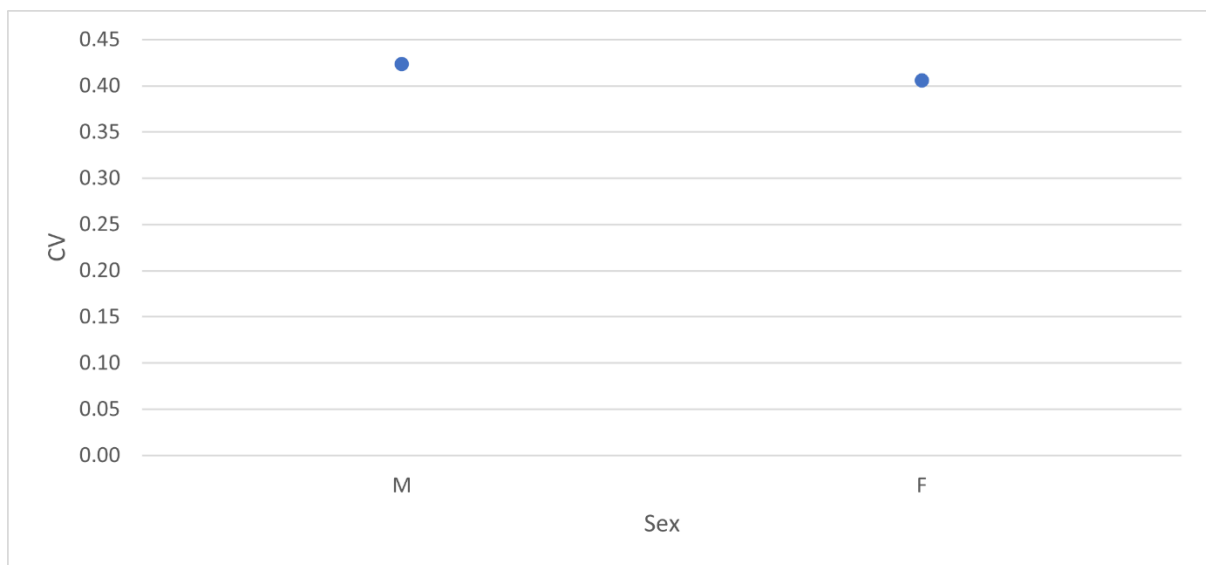


Figure 19.2 CVT by sex for PHQ-9

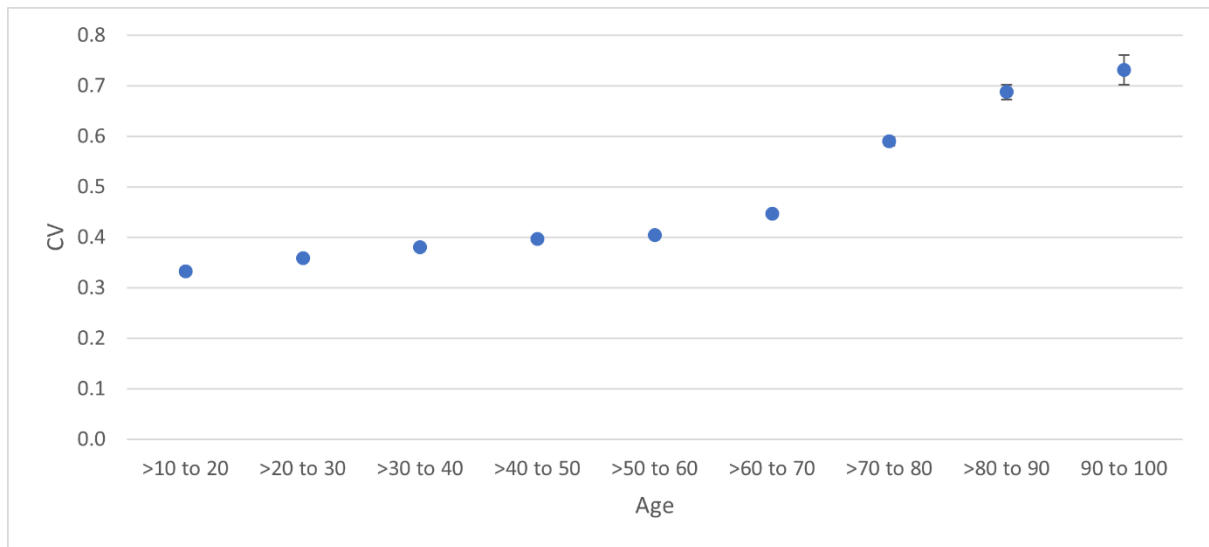


Figure 19.3 CVT by age for PHQ-9

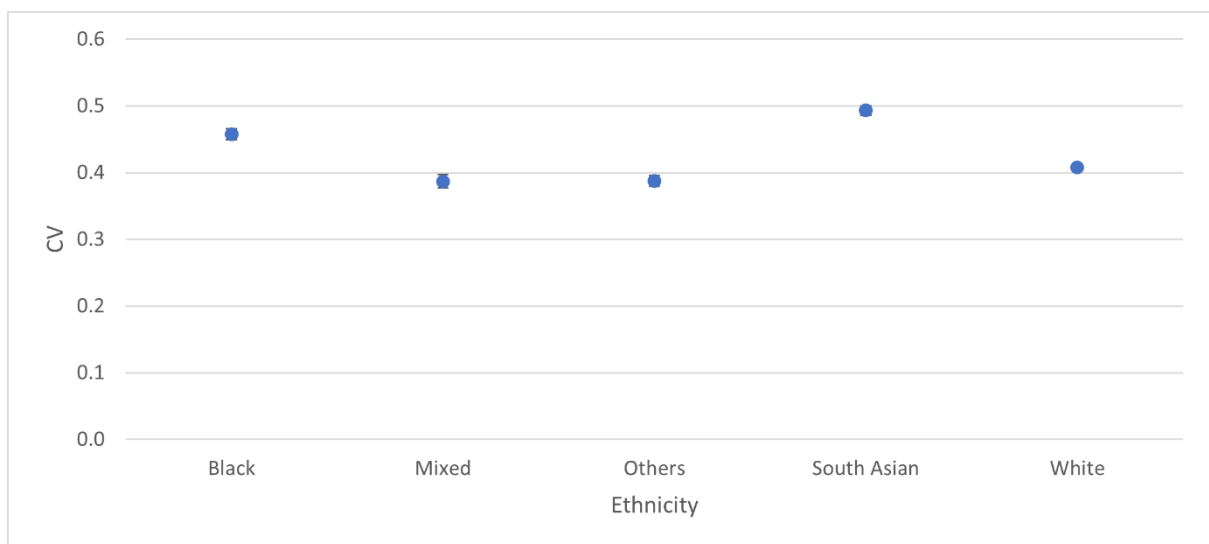


Figure 19.4 CVT by ethnicity for PHQ-9

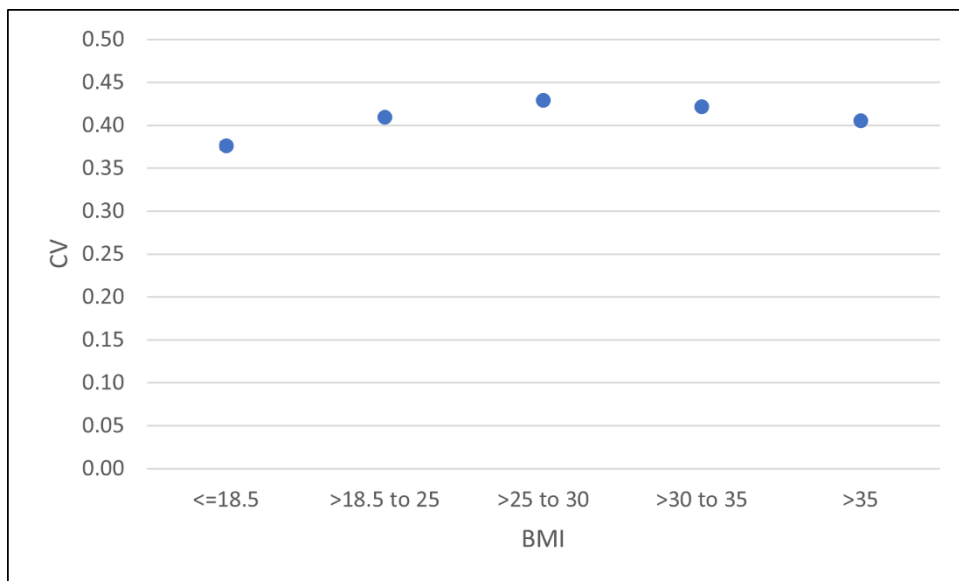


Figure 19.5 CVT by BMI for PHQ-9

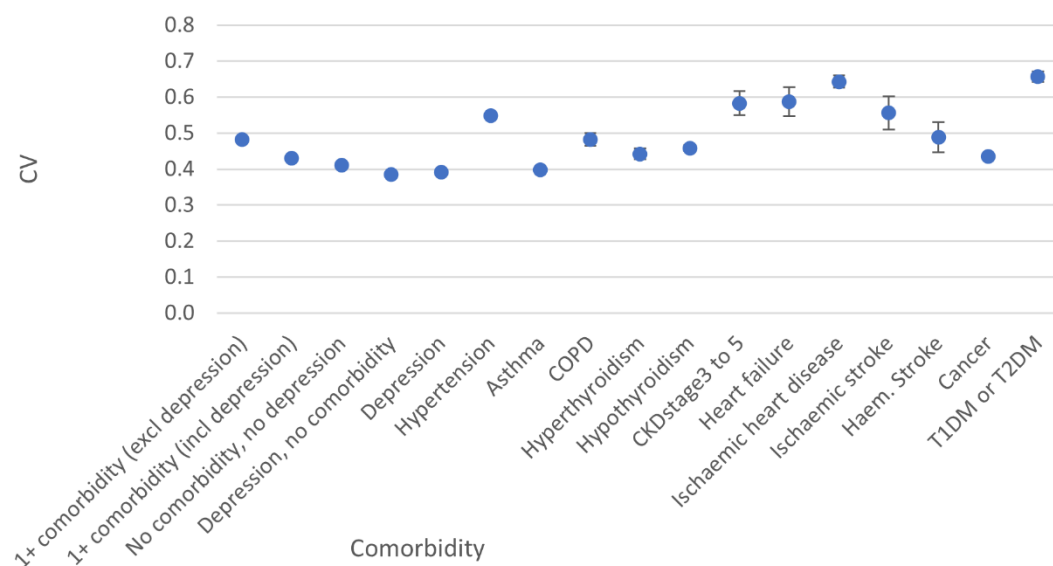


Figure 19.6 CVT by comorbidities for PHQ-9.

First two columns show CV for participants with at least one comorbidity not including depression, and at least one comorbidity including depression.

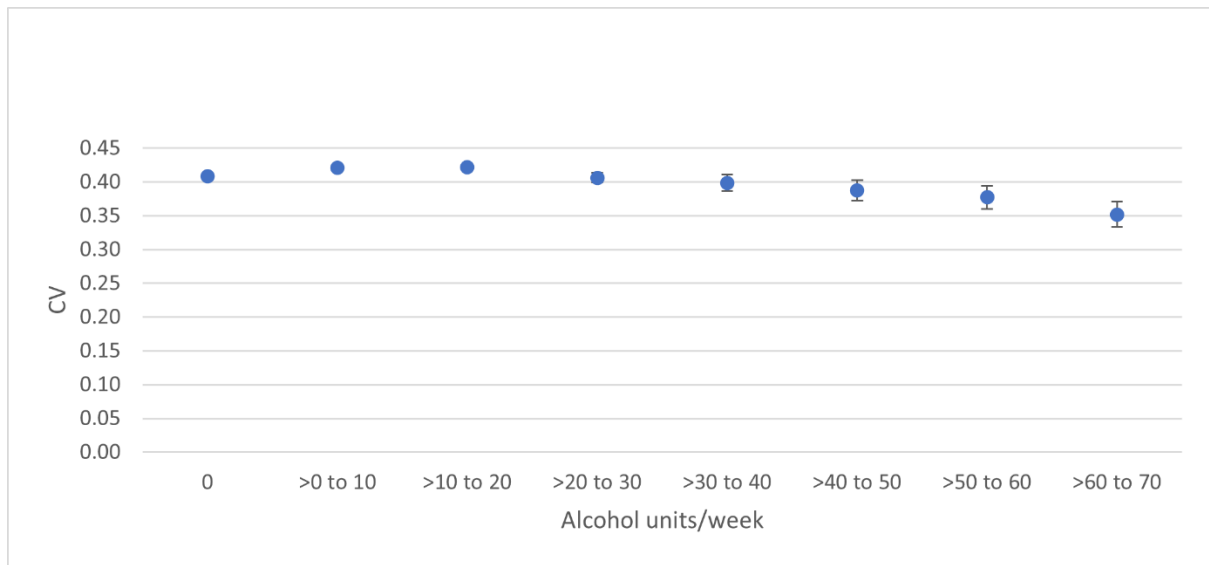


Figure 19.7 CVT by alcohol consumption for PHQ-9

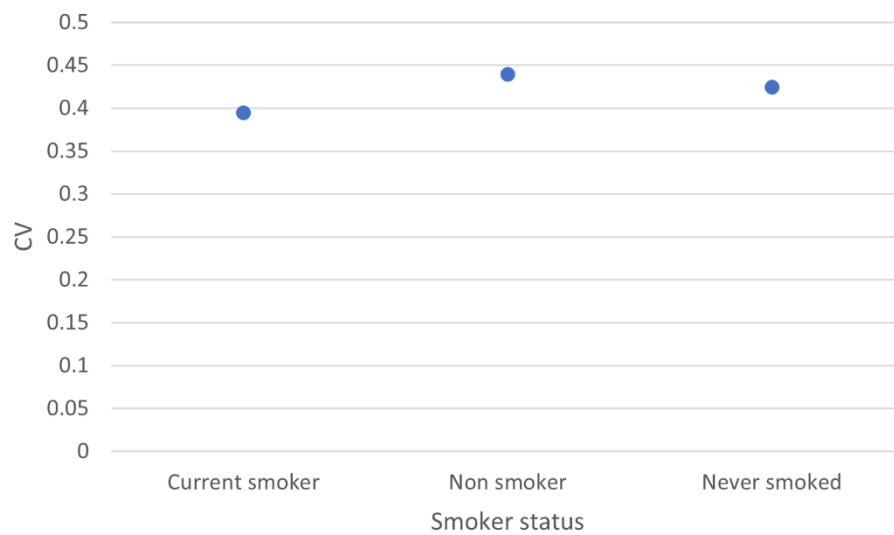


Figure 19.8 CVT by smoker status for PHQ-9

19.2 Appendix 9.2 – sensitivity analyses

Table 19.2 Sensitivity analyses.

*=range

		N	%	Mean	CV (95%CI)	95% LCI	95% UCI
Year	2010	8,517	4.36	13.15	0.368	0.364	0.372
	2011	8,998	4.61	13.26	0.363	0.359	0.366
	2012	9,305	4.77	13.47	0.354	0.350	0.357
	2013	5,341	2.74	13.48	0.338	0.333	0.343
	2014	3,287	1.68	12.96	0.344	0.338	0.351
	2015	2,881	1.48	13.13	0.344	0.337	0.350
	2016	2,294	1.18	13.21	0.337	0.330	0.344
	2017	1,802	0.92	13.26	0.332	0.324	0.340
	2018	1,826	0.94	13.51	0.339	0.331	0.347
	2019	1,974	1.01	13.52	0.327	0.320	0.335
	2010	1,115	0.57	13.41	0.321	0.312	0.331
Number of measurements	2	457,904	N/A	13.15	0.368	0.364	0.372
	3	191,496	N/A	13.26	0.363	0.359	0.366
	4	88,425	N/A	13.47	0.354	0.350	0.357
	5	44,711	N/A	13.48	0.338	0.333	0.343
	6	24,187	N/A	12.96	0.344	0.338	0.351
	7	13,519	N/A	13.13	0.344	0.337	0.350
	8	8,157	N/A	13.21	0.337	0.330	0.344
	9	4,885	N/A	13.26	0.332	0.324	0.340
	10 to 15	8,892	N/A	13.51	0.339	0.331	0.347
	16 to 20	1,429	N/A	13.52	0.327	0.320	0.335
	21 to 25	512	N/A	13.41	0.321	0.312	0.331
	26 to 30	215	N/A	13.15	0.368	0.364	0.372
	31 to 40	147	N/A	13.26	0.363	0.359	0.366
	41 to 50	56	N/A	13.47	0.354	0.350	0.357
Method	Linear regression	12.7		12.69	0.412	0.411	0.413
	Arithmetic	12.66	-	12.66	0.489	0*	4*
Quarter	Q1	14,157	N/A	12.70	0.377	0.373	0.380
	Q2	11,510	N/A	12.48	0.374	0.370	0.378
	Q3	12,133	N/A	12.40	0.384	0.380	0.388
	Q4	12,962	N/A	12.47	0.377	0.373	0.381

	Depression score	Age	Interval between measures				
One or more comorbidity				Y	Y	N	N
Depression diagnosis				Y	N	Y	N
	0 to 4	20 to 60	< 180	1.440	1.414	1.320	1.282
		>60		1.232	1.518	1.244	1.412
		20 to 60	> 180	1.890	2.370	1.309	2.072
		>60		1.886	2.036	1.885	2.066
	5 to 9	20 to 60	< 180	0.658	0.658	0.676	0.652
		>60		0.660	0.678	0.696	0.666
		20 to 60	> 180	0.764	0.785	0.705	0.707
		>60		0.779	0.811	0.757	0.765
	10 to 14	20 to 60	< 180	0.440	0.439	0.445	0.436
		>60		0.449	0.440	0.443	0.435
		20 to 60	> 180	0.480	0.476	0.479	0.464
		>60		0.479	0.492	0.480	0.470
	15 to 19	20 to 60	< 180	0.307	0.292	0.299	0.294
		>60		0.306	0.303	0.306	0.295
		20 to 60	> 180	0.311	0.310	0.314	0.304
		>60		0.312	0.319	0.315	0.320
	20 to 27	20 to 60	< 180	0.169	0.172	0.174	0.171
		>60		0.173	0.188	0.176	0.171
		20 to 60	> 180	0.176	0.172	0.177	0.174
		>60		0.188	0.179	0.172	0.184

Figure 19.9 Heat map including time between measurements for PHQ-9, interval between measures in days

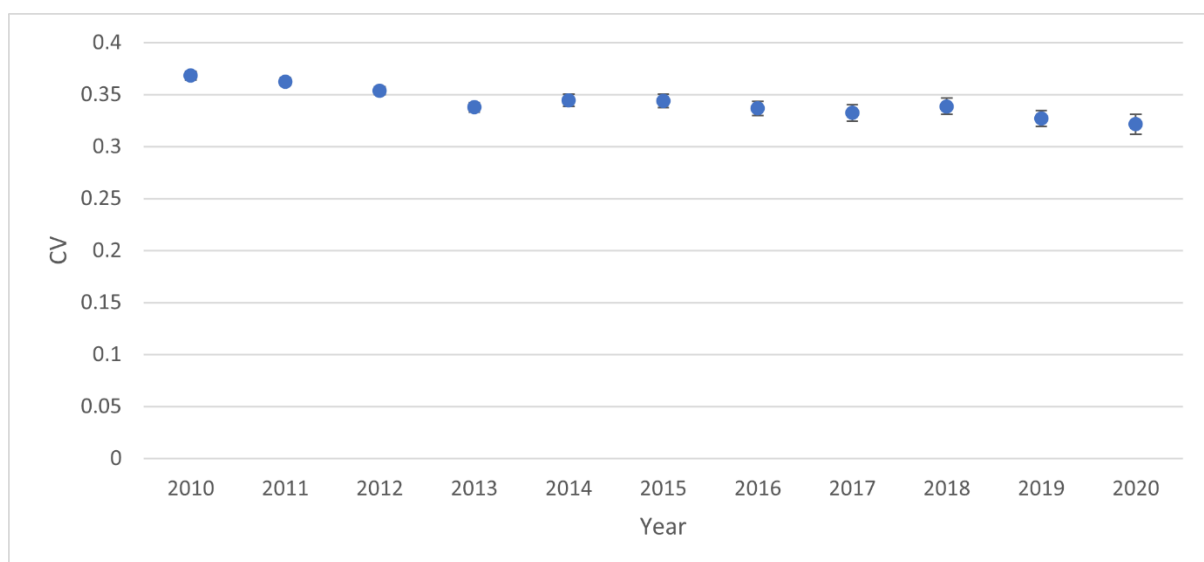


Figure 19.10 CVT by year for PHQ-9

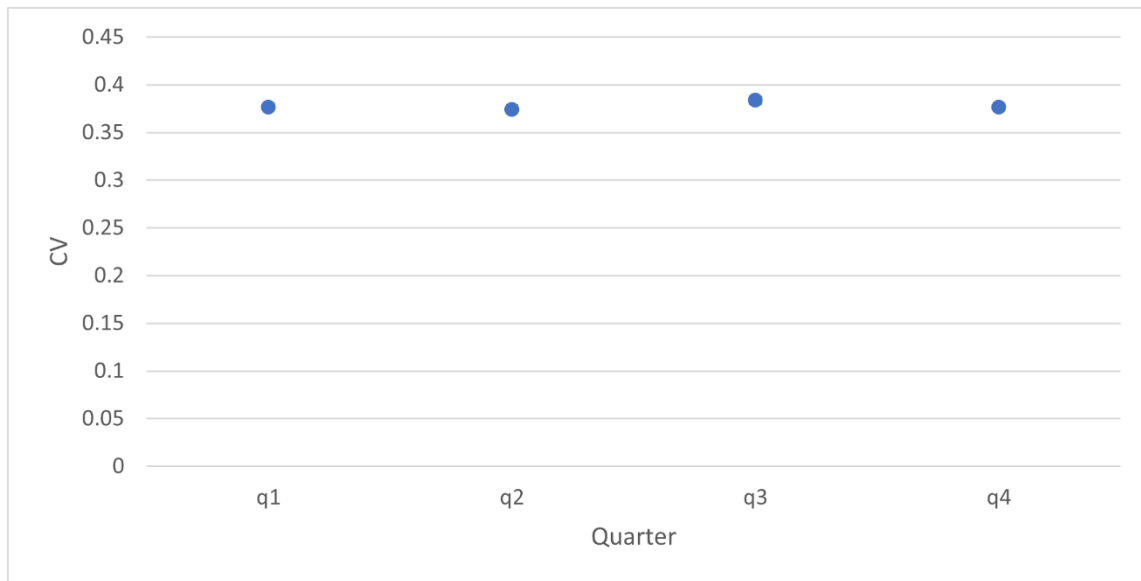


Figure 19.11 CVT by quarter for PHQ-9

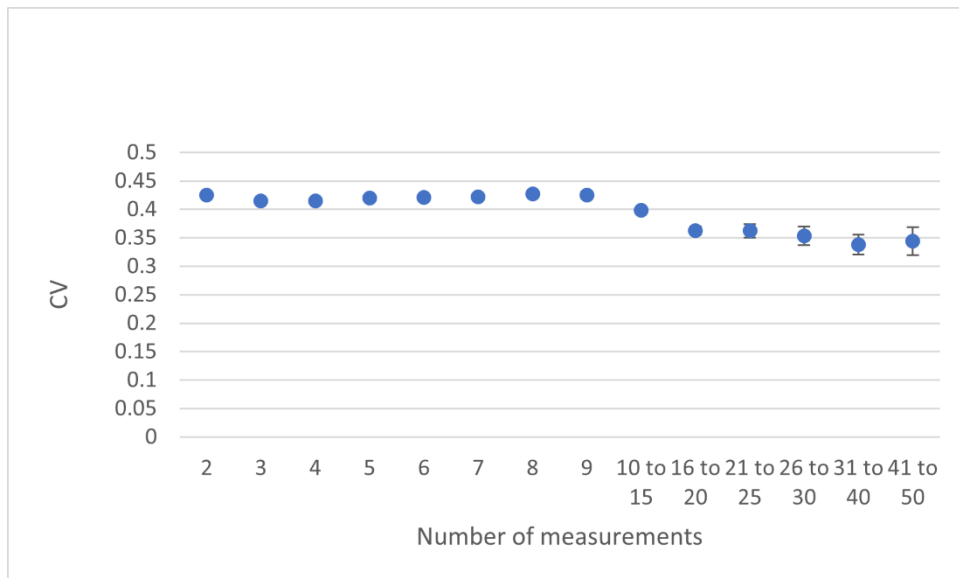


Figure 19.12 CVT by Number of measurements for PHQ-9

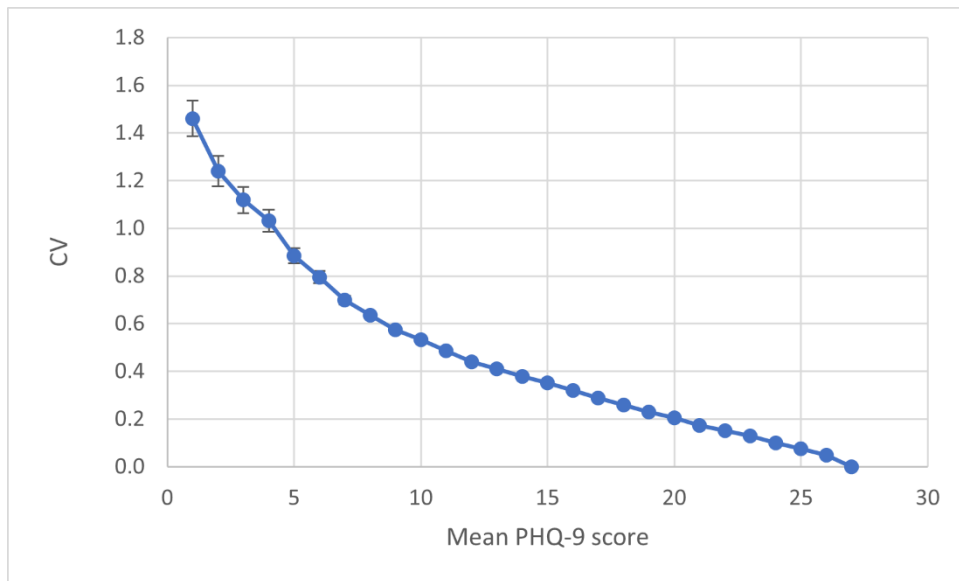


Figure 19.13 CVT by Mean for PHQ-9

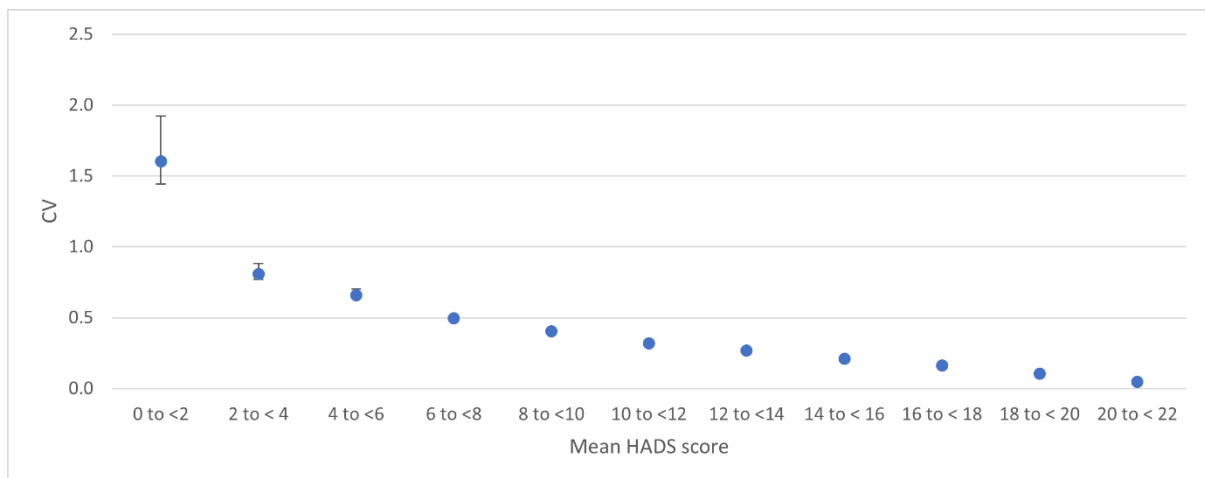


Figure 19.14 CVT by mean patient score for HADS

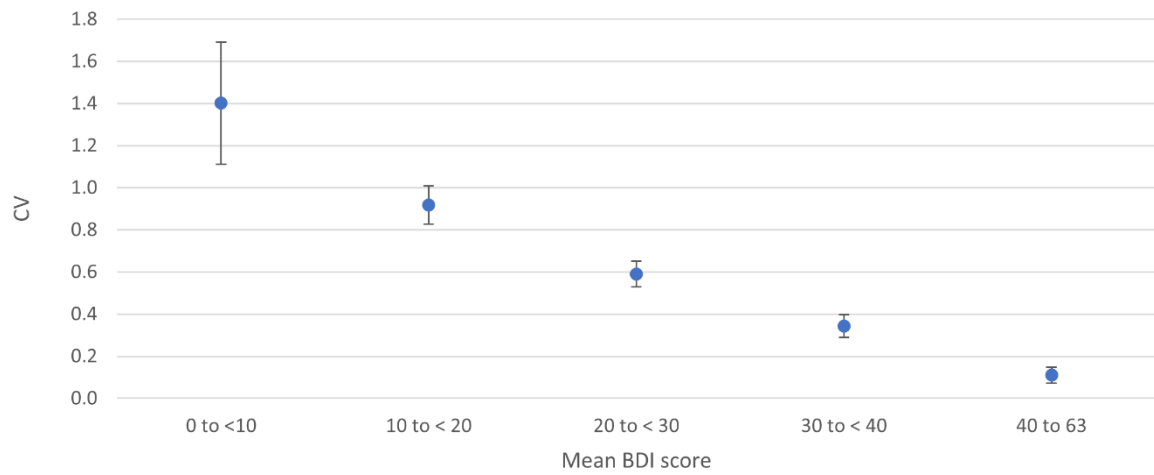


Figure 19.15 CVT by mean patient score for BDI

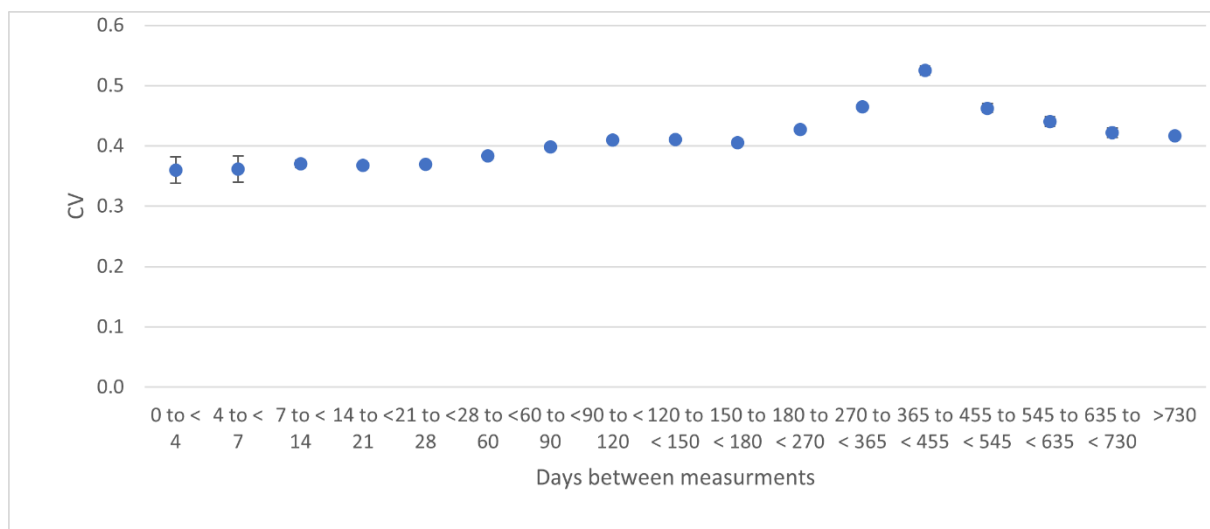


Figure 19.16 CVT by days between measurements for PHQ-9

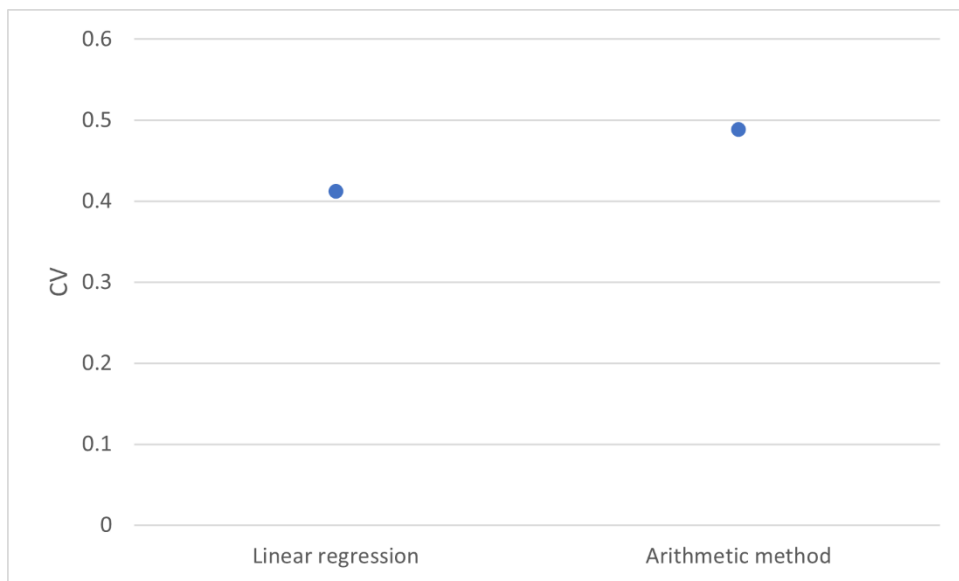


Figure 19.17 CVT by arithmetic calculation methods for PHQ-9

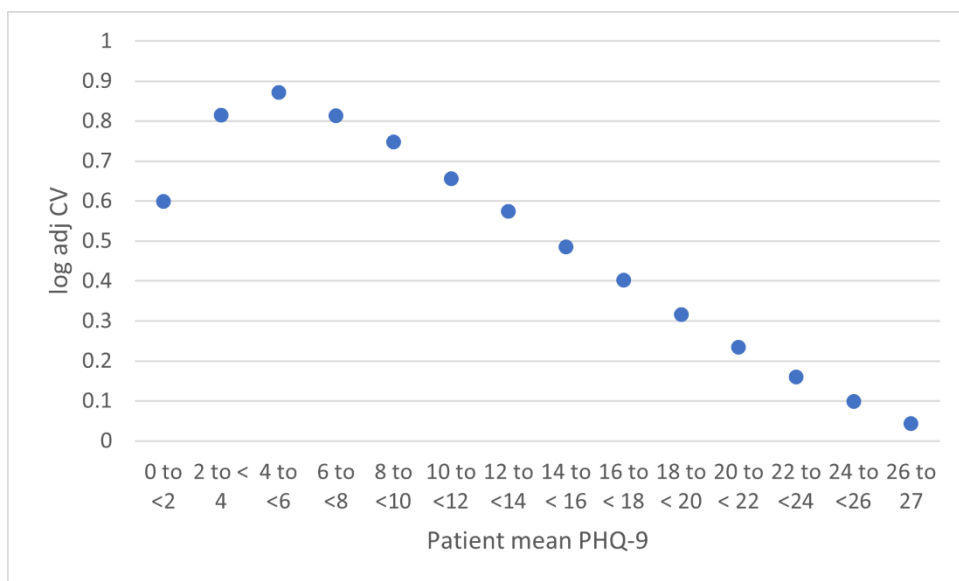


Figure 19.18 CVT by log-transformed data for PHQ-9

20 Appendix 10 – Miscellaneous

20.1 Appendix 10.1 Rapid Scoping Reviews, methods and results

20.1.1 Albumin:creatinine ratio scoping review

Review question:

What is the available data in the literature on biological variation of ACR?

Search strategy:

Search Medline 1946-2019

1. biological varia\$.mp.
2. cvi.mp.
3. within individual.mp.
4. intra-individual.mp.
5. visit to visit.mp.
6. "Reproducibility of Results"/ or coefficient of variation.mp.
7. albumin creatinine ratio.mp.
8. acr.mp.
9. "albumin/creatinine".mp.
10. ratio of albumin to creatinine.mp.
11. albumin-creatinine.mp.
12. 1 or 2 or 3 or 4 or 5 or 6
13. 7 or 8 or 9 or 10 or 11
14. 12 and 13

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability with the coefficient of variation described or able to be calculated of ACR

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Abstracts will be initially screened manually and those with a low probability of relevant data being included in the main text will be excluded

The full text (where found) will then be searched manually for relevant data

Results:

185 papers found. Of these, 70 were excluded as these papers had a different definition of ACR, for example American College of Radiology or area change ratio, 3 as they were on non-human patients, and 101 as they had no data on CvI of ACR.

11 papers had data on CvI or variance of ACR and were included. 9 of these papers had full text available. 10 of the papers had primary data

Reported C_{vi}s for ACR in healthy patients ranged from 11.3% to 31%. In unhealthy patients the reported C_{vi}s of ACR ranged from 13% to 42%.

Discussion

There are a small number of reports of the intra-individual variability of albumin:creatinine ratio, in both healthy and unhealthy individuals. The results of these studies show marked intraindividual variability, and are themselves highly variable. Gibb 1989 states that “within-individual observations for both normals and diabetics are highly but equally variable. Furthermore, from these data, it is possible to infer that a minimum of five estimations are necessary per individual to estimate the true mean value of urine albumin excretion with reasonable confidence.”

However, this is an old paper and only available in abstract. Hatlen noted a higher C_{vi} for women than men. Waikar 2018 notes: “according to present guidelines, an increase in UACR from the 30–300 mg/g range to >300 mg/g denotes a transition from moderately to severely increased albuminuria, with consequences in clinical decision making. For random UACR (median CV_w = 29.7%), an increase from baseline of 124% or a decrease of 55% can be expected by chance.” This study also noted that the degree of variability between individuals was itself highly variable.

In summary, urinary ACR appears to be highly variable within individuals, and the degree of variability is itself highly variable.

20.1.2 CRP scoping review

Rapid scoping review CRP

Review question:

What is the available data in the literature on biological variation of CRP?

Search strategy:

Search Medline 1946-2019

1. biological varia\$.mp.

2. intra-individual varia\$.mp.

3. within individual varia\$.mp.

4. cvi.mp.

5. central venous insufficiency.mp.

6. coefficient of variation.mp.

7. visit to visit.mp.

8. day to day.mp.

9. C-Reactive Protein/ or crp.mp.

10. 4 not 5

11. 1 or 2 or 3 or 10 or 6 or 7 or 8

12. 9 and 11

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability with the coefficient of variation described

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Study is before and after an intervention or a disease

Case reports

Non-human

Abstracts will be initially screened manually and those with a low probability of relevant data being included in the main text will be excluded

The full text (where found) will then be searched manually for relevant data

Results:

249 studies were identified using the search criteria, and after screening, 27 papers with data on variance of CRP remained. 23 of these had primary data. For analysis of this data, where a range of values for Cvl was provided eg for each sex, the mean was taken. One data point was excluded as this used the geometric mean of the Cvl's to provide an outcome. Where variability was expressed other than Cvl eg as variance it was also excluded from the analysis. This left 19 estimations of Cvl from primary sources, which had a mean value of 44% with a SD of 14.5 and a range of 22-74.4%.

Study populations were mainly healthy adults, but some of the studies were conducted on patients undergoing haemodialysis, and others with cystic fibrosis and COPD. The samples sizes ranged from 10 to 300.

Discussion

CRP appears to have a high intraindividual biological variability, although this is not necessarily acknowledged as a problem. One paper concluded, “Our results demonstrate considerable stability and good reproducibility for serial hsCRP and IL-6 measurements. Thus, there should be no major concern about misclassification in clinical practice if at least 2 subsequent measurements are taken.”

There is an increasing recommendation for the use of CRP in guiding antibiotic prescribing, but with as yet little evidence. Although it is noted that there are no prospective trials to recommend this practice yet, it does not seem to be commonly discussed how the variability of CRP may impact on this decision making eg

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5819954/>

CRP is also used in assessment of myocardial disease risk.

Conclusion

A study of CRP may be useful to help ascertain the true biological variability, since both the variability and the range of estimates of variability are large. CRP is being increasingly used in practice for decision making, so may be a useful parameter to study further.

20.1.3 Dementia scoping review

Review question:

What is the available data in the literature on biological variation of tests for Dementia performed in general practice?

Tests recommended by NICE for Dementia include:

the 10-point cognitive screener (10-CS)

the 6-item cognitive impairment test (6CIT)

the 6-item screener

the Memory Impairment Screen (MIS)

the Mini-Cog

Test Your Memory (TYM).

Search strategy:

Search Medline 1946-2020

1. biological variation.mp.

2. cvi.mp.

3. coefficient of variation.mp.

4. within individual.mp.

5. intra-individual.mp.

6. visit to visit.mp.
7. day to day.mp.
8. reproducibility of results.mp. or "Reproducibility of Results"/
9. 10-point cognitive screener.mp.
10. 6-item cognitive impairment test.mp.
11. Memory Impairment Screen.mp.
12. 6-item screener.mp.
13. mini-cog.mp.
14. test your memory.mp.
15. "Mental Status and Dementia Tests"/
16. 9 or 10 or 11 or 12 or 13 or 14 or 15
17. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8
18. 16 and 17

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability with the coefficient of variation described or able to be calculated of one or more of the cognitive assessments

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Abstracts will be initially screened manually and those with a low probability of relevant data being included in the main text will be excluded

The full text (where found) will then be searched manually for relevant data

Results:

The search strategy identified 141 papers. However, examination of these papers showed that where there was any longitudinal data, it mainly involved assessment of cognitive decline or test-retest data to assess for test validity, rather than data on biological variability. Only three papers were identified with any useful data on variability. Feeney et al 2016 examined the MMSE and MoCA and concluded there was considerable intra-patient variability. This limited the diagnostic utility. "An individual's score would have to change by greater than or equal to 3 points on the MMSE and 4 points on the MoCA for the rater to be confident that the change was not due to measurement error." Krishnan et al 2017 concluded that the cognitively intact controls did not change significantly in scores over 3.5 years when measured by MoCA. With a mean change of -0.62 points and a SD of 2.29, Cvl=369% [is this right?!]. Lu et al 2018 aimed to investigate intraindividual variation in normal ageing and vascular disease. They found that the Cvl of reaction times was correlated with MoCA score. They concluded that intraindividual variation of reaction times "is not only associated with poor performance on MoCA subitems but also contributes to their diagnostic accuracy."

Discussion:

Some test-retest and longitudinal studies were identified but no data was available on the intraindividual biological variation of subjects scores on the questionnaire tests of cognitive function. It is possible CVIs could be estimated from the test-retest results if the primary patient data was examined in some of the studies, although only two data points would be available for each individual.

20.1.4 Depression scoping review

Review question:

What is the available data in the literature on biological variation of tests for Depression performed in general practice?

Screening tests recommended for depression include:

Hospital Anxiety and Depression Score

Postnatal depression scale

Geriatric depression scale-15

Patient health Questionnaire-9

Patient health questionnaire-2

Becks depression score

Note that NICE guidelines recommend diagnosis is made on the basis of DSM-IV criteria

Search strategy:

Search Medline 1946-2020

1. biological varia\$.mp.

2. cvi.mp.

3. coefficient of variation.mp.

4. within individual.mp.

5. intra-individual.mp.

6. visit to visit.mp.

7. day to day.mp.

8. (Hospital Anxiety and Depression Score).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

9. postnatal depression scale.mp.

10. Geriatric depression scale-15.mp.

11. Patient health Questionnaire-9.mp. or exp Patient Health Questionnaire/

12. patient health questionnaire-2.mp.

13. becks depression score.mp.

14. 1 or 2 or 3 or 4 or 5 or 6 or 7

15. 8 or 9 or 10 or 11 or 12 or 13

16. 14 and 15

Note – compared to dementia review the search term reproducibility of results was omitted. This seemed to bring up a lot of internal validity tests rather than data on variance. Omitting this search term reduced the number of identified papers from 374 to 7!

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability with the coefficient of variation described or able to be calculated of one or more of the depression assessments

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Abstracts will be initially screened manually and those with a low probability of relevant data being included in the main text will be excluded

The full text (where found) will then be searched manually for relevant data

Results:

Only one of the seven papers had any data related to variance. This was a highly atypical population of patients undergoing haemopoietic cell transplants treated with vorinostat, matched against controls who were also undergoing haemopoietic cell transplants. Although they did not investigate variability specifically, data was included in the paper that should allow a CVI for the different groups to be calculated.

Discussion:

This search found almost no information on variability of depression questionnaires. It is possible that more data might be identified by broadening the search terms, eg including “depression” rather than the depression tests, and including “reproducibility of results” but this is likely to identify a large number of patients with no relevant data.

20.1.5 HbA1c Scoping review

Review question:

What is the available data in the literature on biological variation of HbA1c?

Search strategy:

Database: Ovid MEDLINE(R) <1946 to December Week 4 2019>

Search Strategy:

- 1 Glycated Hemoglobin A/ or hba1c.mp. (45246)
- 2 biological varia\$.mp. (5173)
- 3 intra-individual varia\$.mp. (2013)
- 4 within individual varia\$.mp. (302)
- 5 2 or 3 or 4 (7391)
- 6 1 and 5 (103)

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability with the coefficient of variation described or able to be calculated

Exclusion criteria

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract, or if data only available on inter-individual variability or analytical variability

Results:

109 studies identified, 95 of which had full text available

x studies excluded for lack of relevant data, y included

Of these 109 studies, 26 included data of intraindividual coefficient of variation (CvI) or data to calculate CvI. 6 others were questionable and could be reassessed at a later date. All the rest were excluded for lack of data on CvI except one which was excluded for lack of English translation.

Of these 26 studies included, 12 reported primary data and 14 reported CvI from existing literature. Of those reporting secondary data, some reported data from studies already included in this scoping review, some reported data not included in this scoping review.

For the 12 studies reporting primary data, the crude mean CvI was 3.83%, SD 3.11, range 0.12-8.8

Sample size ranged from 10 to 5952, median 25, IQR 12-71.5

Conclusion/Discussion

Comments and notes from the papers studied included the following observations:

-HbA1c has a marked individuality and the reference ranges are of limited use. Serial measurements are therefore helpful.

- Collecting samples at specific hours of the day or times of the year may reduce high biological variation. Alternatively, the number of individuals may be increased and a paired study design chosen to obtain adequate statistical power.

- Patients with DM showed slightly higher CVI than healthy subjects, this has been seen in previous studies. However, in the only A graded study that included both healthy individuals and stable DM patients, CVI estimates were comparable. Studies reported in mmol/mol should have higher CVs than those reporting % due to differences in methodology, but actually there is little difference.

- HbA1c levels stay remarkably stable over a long period of time

- a small change in Hb A1c may indicate a clinically significant change, e.g., development of diabetes in a patient, even if the difference between the 2 results does not cross the threshold. The CVwp of stable diabetic patients should especially be used for monitoring purposes. However, as shown by Carlsen et al. (3), the CVwp of stable diabetic patients does not differ significantly from that of healthy individuals, and therefore it is probable that the findings in this study will also refer to stable diabetes patients.

- Cvi varies from person to person

- Cvl is different in males and females
- an absolute change smaller than 8.5%, equivalent to 4 HbA1c units at a 50 mmol/mol Hgb, is within uncertainty and should not be assigned clinical significance.
- The number of specimens required to estimate the homeostatic set point within $\pm 6\%$ was 3
- One study suggested that the Cvl itself was correlated with mortality risk
- Some studies examined whether doctors take into account variability during decision making
- There was a controversy over whether use of haemoglobin glycation index (HGI) is a useful and statistically valid parameter

20.1.6 Lipids scoping review

Review question:

What is the available data in the literature on biological variation of lipids:

Cholesterol, LDL, HDL, VLDL and triglycerides?

Search strategy:

Search Medline 1946-2019

1. biological varia\$.mp.

2. intra-individual varia\$.mp.

3. within individual varia\$.mp.
4. biological varia\$.mp.
5. cvi.mp.
6. within individual.mp.
7. intra-individual.mp.
8. visit to visit.mp.
9. coefficient of variation.mp.
10. day to day.mp.
11. Cholesterol, LDL/ or Cholesterol, VLDL/ or Cholesterol/ or Cholesterol, HDL/ or cholesterol.mp.
12. triglycerides.mp. or exp Triglycerides/
13. low density lipoprotein.mp. or Lipoproteins, LDL/
14. very low density lipoprotein.mp. or Lipoproteins, VLDL/
15. high density lipoprotein.mp. or Lipoproteins, HDL/
16. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10
17. 11 or 12 or 13 or 14 or 15
18. 16 and 17

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability with the coefficient of variation described or able to be calculated of one of the lipid biomarkers

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Abstracts will be initially screened manually and those with a low probability of relevant data being included in the main text will be excluded

The full text (where found) will then be searched manually for relevant data

Results:

967 papers were identified in the initial search. 782 papers were excluded after screening abstract and title for being unlikely to have data related to the review and 2 papers for lack of abstract or full text leaving 183 papers. Full texts were sought for each of these.

(how did you get from 183 to 81?)

81 papers met the eligibility criteria for the review, having data on CvI of one or more of the lipids, or in some cases another measure of intra-individual variability.

For numerical analysis, only data with a single value for CvI was included.

For total cholesterol, the mean CvI was 6.1% out of 64 data points

For LDL, 9.34% from 71 data points

For HDL, 7.03% from 71 data points

For Trig, 21.46% from 71 data points

Discussion

For decades the of high intraindividual variability in lipid measurements, particularly triglycerides has been discussed. Causes of systematic variation include seasonal, diurnal, and fasting status, but there was also a high level of random variation even after controlling for these. The need for multiple measurements is well recognised. One study suggested that, “that six tests have to be taken before and after intervention by drug or diet to detect a reduction of 10% in an individual with a probability of 0.95.” More recent papers have focussed on variability of lipids as an independent risk factor, eg for CV events, coronary atheroma progression, end stage renal disease, dementia.

Conclusion

There is a large body of work on the biological variation of lipids, and the problems associated with this are well recognised. Lipids are therefore not likely to be good candidates for further study.

20.1.7 Spirometry scoping review

Review question:

What is the available data in the literature on biological variation of spirometry measures FEV1 and FVC?

Search strategy:

Medline 1946-2019

1. spirometry.mp. or Spirometry/
2. Fev.mp. or Forced Expiratory Volume/
3. FVC1.mp.
4. FVC.mp. or Respiratory Function Tests/ or Vital Capacity/
5. biological variability.mp.
6. biological variation.mp.
7. within-individual variation.mp.
8. intra-individual variation.mp.
9. 5 or 6 or 7 or 8
10. 1 or 2 or 3 or 4
11. 9 and 10

Inclusion criteria

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability, inter-individual biological variability, analytical variability, with the coefficient of variation described or able to be calculated

Exclusion criteria

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Results

18 studies were identified of which 15 had full text available. 15 studies were excluded for absence of relevant data. Of the remaining 3 studies, only one provided data on Cvl, and this was a 1989 study of 232 7 year old children. One other study which was excluded since it did not provide information about Cvl gave information on the variance (sigma squared) of repeated measures of FEV1 in an individual, but the Cvl did not appear to be calculable from the data provided.

Conclusion

There is very little data on the intra-individual variability of spirometry tests.

20.1.8 TSH scoping review

Review question:

What is the available data in the literature on biological variation of TSH?

Search strategy:

Search Medline 1946-2019

1. biological variation\$.mp.
2. Thyrotropin/ or TSH.mp.
3. thyroid stimulating hormone.mp. or Thyrotropin/
4. 2 and 3
5. 1 and 4

Iterative searching not performed in this review

Inclusion criteria:

Studies will be included if:

Published in English language, or translation available

Study provides data on the intra-individual biological variability, inter-individual biological variability, analytical variability, with the coefficient of variation described or able to be calculated

Studies will be excluded if:

Full text not available, and pertinent data is not included in the abstract

Results:

25 studies identified, 21 of which had full text available

8 studies excluded for lack of relevant data, 17 included

Of these 17 studies, 7 included data of intraindividual coefficient of variation (CvI) or data to calculate CvI

Of these 7 studies, 4 reported primary data and 3 reported CvI from existing literature

Only three studies of adults were available with a total of 48 patients (but more studies may be located if full text can be found and iterative searching is performed).

CvI of TSH in healthy adults varied from 16.2% to 29.3%.

Discussion

There are few primary studies reporting CvI of TSH, and those studies are of very few patients. Some points of interest were noted while performing the literature search. Kouri et al noted that the measurement uncertainty of cholesterol, free thyroxine, TSH and C-reactive protein were mainly due to biological variability. Maes et al 1997 noted that it might be difficult to compare data from older papers eg from 1985 as they aren't using more modern ultra-high sensitivity tests for TSH. This paper also notes that TSH shows a marked seasonal variation. Padro-Miguel noted that in general there is a big variation in the results of variability studies, and that biological variability as a way of assessing an individual may not be helpful (they propose a different method). The intra-individual variability may vary greatly between individuals (distinct from the inter-individual variability). There may therefore be a benefit in calculating not only a mean co-efficient of variability for TSH and other parameters, but to calculate a between individual coefficient of variability of the within individual coefficient of variability ie a CvG of the CvI.

21 References

1. McCormack JP, Holmes DT. Your results may vary: the imprecision of medical measurements. *BMJ*. 2020;368:m149.
2. Ricos C, Iglesias N, Garcia-Lario JV, Simon M, Cava F, Hernandez A, et al. Within-subject biological variation in disease: collated data and clinical consequences. *Ann Clin Biochem*. 2007;44(Pt 4):343-52.
3. Anonymous. Hypertension in adults: diagnosis and management, NICE guideline [NG136] 2019 [Available from: <https://www.nice.org.uk/guidance/ng136/>].
4. Wallis A, Edey A, Prothero D, McCoubrie P. The Bristol Radiology Report Assessment Tool (BRRAT): Developing a workplace-based assessment tool for radiology reporting skills. *Clinical Radiology*. 2013;68(11):1146-54.
5. Glover GH. Overview of functional magnetic resonance imaging. *Neurosurg Clin N Am*. 2011;22(2):133-9, vii.
6. Mannil M, Polysopoulos C, Weishaupt D, Hansmann A. Clinical-radiological scoring system for enhanced diagnosis of acute appendicitis. *Eur J Radiol*. 2018;98:174-8.
7. Cleveland RH, Stamoulis C, Sawicki G, Kelliher E, Zucker EJ, Wood C, et al. Brasfield and Wisconsin scoring systems have equal value as outcome assessment tools of cystic fibrosis lung disease. *Pediatr Radiol*. 2014;44(5):529-34.
8. López-Alfaro MP, Echarte-Nuin I, Fernández-Sangil P, Moyano-Berardo BM, Goñi-Viguria R. Perception of pain in post-surgical patients in intensive care units. *Enferm Intensiva*. 2019;30(3):99-107.
9. Pucher PH, Qurashi M, Howell AM, Faiz O, Ziprin P, Darzi A, et al. Development and validation of a symptom-based severity score for haemorrhoidal disease: the Sodergren score. *Colorectal Dis*. 2015;17(7):612-8.
10. Feeney J, Savva GM, O'Regan C, King-Kallimanis B, Cronin H, Kenny RA. Measurement Error, Reliability, and Minimum Detectable Change in the Mini-Mental State Examination, Montreal Cognitive Assessment, and Color Trails Test among Community Living Middle-Aged and Older Adults. *Journal of Alzheimer's Disease*. 2016;53(3):1107-14.
11. Hoodin F, LaLonde L, Errickson J, Votruba K, Kentor R, Gatza E, et al. Cognitive Function and Quality of Life in Vorinostat-Treated Patients after Matched Unrelated Donor Myeloablative Conditioning Hematopoietic Cell Transplantation. *Biology of Blood & Marrow Transplantation*. 2019;25(2):343-53.
12. Whiting P, Toerien M, de Salis I, Sterne JAC, Dieppe P, Egger M, et al. A review identifies and classifies reasons for ordering diagnostic tests. *Journal of Clinical Epidemiology*. 2007;60(10):981-9.
13. van der Weijden T. Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *British Journal of General Practice*. 52(485):974-81.
14. Lara J. A proposed panel of biomarkers of healthy ageing. *BMC Medicine*. 13(1):1-9.
15. Fraser Callum G. Biological Variation: From Principles to Practice. USA: AACC Press; 2001.
16. Livingstone R. How tightly controlled do fluctuations in blood glucose levels need to be to reduce the risk of developing complications in people with Type 1 diabetes? *Diabetic Medicine*. 37(4):513-22.
17. Anonymous. Global Initiative for Asthma (GINA), Pocket Guide for Asthma Management and Prevention, 2019 [Available from: <https://ginasthma.org/wp-content/uploads/2019/04/GINA-2019-main-Pocket-Guide-wms.pdf>].

18. Lima-Oliveira G, Volanski W, Lippi G, Picheth G, Guidi GC. Pre-analytical phase management: a review of the procedures from patient preparation to laboratory analysis. *Scandinavian Journal of Clinical and Laboratory Investigation*. 2017;77(3):153-63.
19. Tolonen H, Ferrario M, Kuulasmaa K. Standardization of total cholesterol measurement in population surveys--pre-analytic sources of variation and their effect on the prevalence of hypercholesterolaemia. *Eur J Cardiovasc Prev Rehabil*. 2005;12(3):257-67.
20. Lippi G, Mattiuzzi C, Favaloro EJ. Pre-analytical variability and quality of diagnostic testing. Looking at the moon and gazing beyond the finger. *New Zealand Journal of Medical Laboratory Science*. 2015;69(1):4-8.
21. Lima-Oliveira G, Guidi GC, Salvagno GL, Danese E, Montagnana M, Lippi G. Patient posture for blood collection by venipuncture: recall for standardization after 28 years. *Revista Brasileira de Hematologia e Hemoterapia*. 39(2):127-32.
22. Fagard HR, Stolarz AK, Kuznetsova AT, Seidlerova AJ, Tikhonoff AV, Grodzicki AT, et al. Sympathetic activity, assessed by power spectral analysis of heart rate variability, in white-coat, masked and sustained hypertension versus true normotension. *Journal of Hypertension*. 25(11):2280-5.
23. Rifkin L, Schaal S. Factors affecting patients' pain intensity during in office intravitreal injection procedure. *Retina*. 2012;32(4):696-700.
24. Dominguez-Rodriguez A, Abreu-Gonzalez P. Myeloperoxidase in the acute coronary syndrome: Equal concentrations at any time of day? *International Journal of Cardiology*. 2011;150(2):206-7.
25. Strobel NA, Fassett RG, Marsh SA, Coombes JS. Importance of understanding pre-analytical variability in biomarker development. *International Journal of Cardiology*. 2011;150(2):223-4.
26. Westgard JO. Useful measures and models for analytical quality management in medical laboratories. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2016;54(2):223-33.
27. Mao X, Shao J, Zhang B, Wang Y. Evaluating analytical quality in clinical biochemistry laboratory using Six Sigma. *Biochem Med (Zagreb)*. 2018;28(2):020904.
28. Irjala KM, Gronroos PE. Preanalytical and analytical factors affecting laboratory results. *Ann Med*. 1998;30(3):267-72.
29. Marlara RA, Rollins-Raval MA. Sources and solutions for spurious test results in coagulation. *Int J Lab Hematol*. 2019;41 Suppl 1:162-9.
30. Westgard S, Petrides V, Schneider S, Berman M, Herzogenrath J, Orzechowski A. Assessing precision, bias and sigma-metrics of 53 measurands of the Alinity ci system. *Clin Biochem*. 2017;50(18):1216-21.
31. Schultze AE, Irizarry AR. Recognizing and Reducing Analytical Errors and Sources of Variation in Clinical Pathology Data in Safety Assessment Studies. *Toxicol Pathol*. 2017;45(2):281-7.
32. Oosterhuis WP, Bayat H, Armbruster D, Coskun A, Freeman KP, Kallner A, et al. The use of error and uncertainty methods in the medical laboratory. *Clin Chem Lab Med*. 2018;56(2):209-19.
33. Schermer T, Schermer TRJ, Verweij EHA, Cretier R, Pellegrino A, Crockett AJ, et al. Accuracy and Precision of Desktop Spirometers in General Practices. *Respiration*. 2012;83(4):344-52.
34. Arndt JL, Pfau T, Day P, Pardoe C, Bolt DM, Weller R. Forces applied with a hoof tester to cadaver feet vary widely between users. *Vet Rec*. 2013;172(7):182.
35. Novoa YNVA, Sutton LF, Neis AE, Marroquin AM, Coleman TM, Praska KA, et al. Reproducibility of Liver-to-Thorax Area Ratio Ultrasound Measurements in Congenital Diaphragmatic Hernia. *J Ultrasound Med*. 2019;38(6):1477-82.

36. Plasencia-Martinez JM, Garcia-Santos JM, Lozano-Herrero J, Hernandez-Vidal MJ. Carotid Intima-Media Thickness Manual Measurements: Intraoperator and Interoperator Agreements Under A Strict Protocol in a Large Sample. *Ultrasound Q.* 2017;33(1):28-36.
37. Wolfe F, Pincus T, Thompson AK, Doyle J. The assessment of rheumatoid arthritis and the acceptability of self-report questionnaires in clinical practice. *Arthritis Care & Research.* 2003;49(1):59-63.
38. Bradshaw AE, McNamara C. 24 - Laboratory Organisation, Management and Safety. In: Bain BJ, Bates I, Laffan MA, editors. *Dacie and Lewis Practical Haematology (Twelfth Edition)*: Elsevier; 2017. p. 511-32.
39. Bodger K, Ormerod C, Shackcloth D, Harrison M, Collaborative IBDC. Development and validation of a rapid, generic measure of disease control from the patient's perspective: the IBD-control questionnaire. *Gut.* 2014;63(7):1092-102.
40. Simundic A-M, Kackov S, Miler M, Fraser CG, Petersen PH. Terms and Symbols Used in Studies on Biological Variation: The Need for Harmonization. *Clinical Chemistry.* 2015;61(2):438-9.
41. Ricos C, Alvarez V, Perich C, Fernandez-Calle P, Minchinela J, Cava F, et al. Rationale for using data on biological variation. *Clin Chem Lab Med.* 2015;53(6):863-70.
42. Hallgrímsson B, Hall BK. *Variation : A Central Concept in Biology.* San Diego, UNITED STATES: Elsevier Science & Technology; 2005.
43. Marshall T. When measurements are misleading: modelling the effects of blood pressure misclassification in the English population. *BMJ.* 2004;328(7445):933.
44. Marshall T. Misleading Measurements: Modeling the Effects of Blood Pressure Misclassification in a United States Population. *Medical Decision Making.* 2006;26(6):624-32.
45. Anonymous. Pneumonia in adults: diagnosis and management. NICE guideline cg191 2014 [cited 2020 21/04/2020]. Available from: <https://www.nice.org.uk/guidance/cg191/resources/pneumonia-in-adults-diagnosis-and-management-pdf-35109868127173>.
46. Malaponte G, Libra M, Bevelacqua Y, Merito P, Fatuzzo P, Rapisarda F, et al. Inflammatory status in patients with chronic renal failure: the role of PTX3 and pro-inflammatory cytokines. *International Journal of Molecular Medicine.* 2007;20(4):471-81.
47. Nunes LA, Brenzikofer R, de Macedo DV. Reference change values of blood analytes from physically active subjects. *European Journal of Applied Physiology.* 2010;110(1):191-8.
48. Neild G. Life expectancy with chronic kidney disease: an educational review. *Pediatric Nephrology.* 32(2):243-9.
49. Anonymous. Chronic kidney disease in adults: assessment and management: NICE guidelines 2014 [Available from: <https://www.guidelines.co.uk/renal-conditions/nice-ckd-guideline-/252515.article>].
50. Selvin E, Juraschek SP, Eckfeldt J, Levey AS, Inker LA, Coresh J. Within-person variability in kidney measures. *American Journal of Kidney Diseases.* 2013;61(5):716-22.
51. Waikar SS, Rebholz CM, Zheng Z, Hurwitz S, Hsu CY, Feldman HI, et al. Biological Variability of Estimated GFR and Albuminuria in CKD. *American Journal of Kidney Diseases.* 2018;72(4):538-46.
52. Kringos DS, Boerma WGW, Hutchinson A, van der Zee J, Groenewegen PP. The breadth of primary care: a systematic literature review of its core dimensions. *BMC Health Services Research.* 2010;10(1):65.
53. Kringos D, Boerma W, Bourgueil Y, Cartier T, Dedeu T, Hasvold T, et al. The strength of primary care in Europe: an international comparative study. *Br J Gen Pract.* 2013;63(616):e742-50.
54. O'Sullivan JW, Stevens S, Hobbs FDR, Salisbury C, Little P, Goldacre B, et al. Temporal trends in use of tests in UK primary care, 2000-15: retrospective analysis of 250 million tests. *Bmj.* 2018;363:k4666.

55. Anonymous. Quality and Outcomes Framework 2020 [Available from: <https://qof.digital.nhs.uk/>].
56. Anonymous. The National Institute for Health and Care Excellence 2020 [Available from: <https://www.nice.org.uk/>].
57. Ceriotti F. Prerequisites for use of common reference intervals. *Clin Biochem Rev*. 2007;28(3):115-21.
58. Katayev A, Balciza C, Seccombe DW. Establishing Reference Intervals for Clinical Laboratory Test Results: Is There a Better Way? *American Journal of Clinical Pathology*. 2010;133(2):180-6.
59. Christensen RD, Henry E, Jopling J, Wiedmeier SE. The CBC: reference ranges for neonates. *Semin Perinatol*. 2009;33(1):3-11.
60. Mugisha JO, Seeley J, Kuper H. Population based haematology reference ranges for old people in rural South-West Uganda. *BMC Res Notes*. 2016;9(1):433.
61. D'Ascenzi F, Pelliccia A, Solari M, Piu P, Loiacono F, Anselmi F, et al. Normative Reference Values of Right Heart in Competitive Athletes: A Systematic Review and Meta-Analysis. *J Am Soc Echocardiogr*. 2017;30(9):845-58.e2.
62. Wener MH, Daum PR, McQuillan GM. The influence of age, sex, and race on the upper reference limit of serum C-reactive protein concentration. *J Rheumatol*. 2000;27(10):2351-9.
63. Jensen E, Hyltoft Petersen P, Blaabjerg O, Hansen PS, Brix TH, Kyvik KO, et al. Establishment of a serum thyroid stimulating hormone (TSH) reference interval in healthy adults. The importance of environmental factors, including thyroid antibodies. *Clin Chem Lab Med*. 2004;42(7):824-32.
64. Anonymous. NICE guidance on hypothyroidism 2020 [cited 2020 8.4.20]. Available from: <https://cks.nice.org.uk/hypothyroidism#!diagnosisSub:1>.
65. Fraser CG. Test result variation and the quality of evidence-based clinical guidelines. *Clinica Chimica Acta*. 2004;346(1):19-24.
66. Pearson AT, Mensah AG, Alexander WR, Anderson LJ, Cannon OR, Criqui YM, et al. Markers of Inflammation and Cardiovascular Disease: Application to Clinical and Public Health Practice: A Statement for Healthcare Professionals From the Centers for Disease Control and Prevention and the American Heart Association. *Circulation: Journal of the American Heart Association*. 107(3):499-511.
67. Tracy CS, Dantas GC, Moineddin R, Upshur REG. Contextual factors in clinical decision making: national survey of Canadian family physicians. *Canadian Family Physician*. 2005;51(8):1106.
68. Browning MC, Ford RP, Callaghan SJ, Fraser CG. Intra- and interindividual biological variation of five analytes used in assessing thyroid function: implications for necessary standards of performance and the interpretation of results. *Clinical Chemistry*. 1986;32(6):962-6.
69. Maes M, Mommen K, Hendrickx D, Peeters D, D'Hondt P, Ranjan R, et al. Components of biological variation, including seasonality, in blood concentrations of TSH, TT3, FT4, PRL, cortisol and testosterone in healthy volunteers. *Clinical Endocrinology*. 1997;46(5):587-98.
70. Hollowell JG, Staehling NW, Flanders WD, Hannon WH, Gunter EW, Spencer CA, et al. Serum TSH, T(4), and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III). *J Clin Endocrinol Metab*. 2002;87(2):489-99.
71. Yousuf O, Mohanty BD, Martin SS, Joshi PH, Blaha MJ, Nasir K, et al. High-sensitivity C-reactive protein and cardiovascular disease: a resolute belief or an elusive link? *Journal of the American College of Cardiology*. 62(5):397-408.
72. Woloshin S, Schwartz LM. Distribution of C-reactive protein values in the United States. *N Engl J Med*. 2005;352(15):1611-3.

73. Llor C, Bjerrum L, Arranz J, Garcia G, Cots JM, Gonzalez Lopez-Valcarcel B, et al. C-reactive protein testing in patients with acute rhinosinusitis leads to a reduction in antibiotic use. *Fam Pract.* 2012;29(6):653-8.
74. Pereira MA, Weggemans RM, Jacobs DR, Jr., Hannan PJ, Zock PL, Ordovas JM, et al. Within-person variation in serum lipids: implications for clinical trials. *International Journal of Epidemiology.* 2004;33(3):534-41.
75. Monsalve MV, Robinson D, Woolcock NE, Powell JT, Greenhalgh RM, Humphries SE. Within-individual variation in serum cholesterol levels: association with DNA polymorphisms at the apolipoprotein B and AI-CIII-AIV loci in patients with peripheral arterial disease. *Clinical Genetics.* 1991;39(4):260-73.
76. Stevens SL, Wood S, Koshiaris C, Law K, Glasziou P, Stevens RJ, et al. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. *BMJ.* 2016;354:i4098.
77. EFLM. EFLM Biological Variation Database. [Available from: <https://biologicalvariation.eu/>].
78. Fraser GG, Harris EK. Generation and Application of Data on Biological Variation in Clinical Chemistry. *Critical Reviews in Clinical Laboratory Sciences.* 1989;27(5):409-37.
79. Vuvor F, Harrison O. A study of the diurnal height changes among sample of adults aged 30 years and above in Ghana. *Biomedical and Biotechnology Research Journal (BBRJ).* 2017;1(2):113-9.
80. Quer G, Gouda P, Galarnyk M, Topol EJ, Steinhubl SR. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *PLoS One.* 2020;15(2):e0227709.
81. Liu JHK, Sit AJ, Weinreb RN. Variation of 24-hour intraocular pressure in healthy individuals: right eye versus left eye. *Ophthalmology.* 112(10):1670.
82. Strachan DP. Repeatability of ventilatory function measurements in a population survey of 7 year old children. *Thorax.* 44(6):474-9.
83. Seliem MA, McWilliams ET, Palileo M. Beat-to-beat variability of left ventricular indexes measured by acoustic quantification: influence of heart rate and respiration--correlation with M-mode echocardiography. *J Am Soc Echocardiogr.* 1996;9(3):221-30.
84. Grimm M, Koziolok M, Kuhn JP, Weitschies W. Interindividual and intraindividual variability of fasted state gastric fluid volume and gastric emptying of water. *Eur J Pharm Biopharm.* 2018;127:309-17.
85. Mohidin N, Lau Chean L. Central and peripheral corneal thickness in Malays and its variation with age. *Bangladesh Journal of Medical Science.* 2018;17(4):600-5.
86. Lee EY, Kim D, Kim KM, Kim KJ, Choi HS, Rhee Y, et al. Age-Related Bone Mineral Density Patterns in Koreans (KNHANES IV). *The Journal of Clinical Endocrinology & Metabolism.* 2012;97(9):3310-8.
87. Alexander H, Cook T. Variations with age in the mechanical properties of human skin in vivo. *Journal of Tissue Viability.* 2006;16(3):6-11.
88. Galbusera F, Rijsbergen M, Ito K, Huyghe J, Brayda-Bruno M, Wilke H-J. Ageing and degenerative changes of the intervertebral disc and their impact on spinal flexibility. *European Spine Journal.* 23(Supplement 3):324-32.
89. Millar AC, Lau ANC, Tomlinson G, Kraguljac A, Simel DL, Detsky AS, et al. Predicting low testosterone in aging men: a systematic review. *CMAJ.* 2016;188(13):E321-E30.
90. Peterson DD. Longitudinal and cross-sectional analysis of HEALTH WATCH data with a subset of perimenopausal women and matched controls. *Journals of Gerontology Series A: Biological Sciences & Medical Sciences.* 54(4):B160.
91. Hishikawa N. Clinical predictors of Alzheimer's disease progression. *Geriatrics & Gerontology International.* 18(6):929-37.

92. Lucke C, Hehrmann R, Von Mayersbach K, Von Zur Mühlen A. STUDIES ON CIRCADIAN VARIATIONS OF PLASMA TSH, THYROXINE AND TRIIODOTHYRONINE IN MAN. *Acta Endocrinologica*.86(1):81-8.
93. Pruessner JC, Wolf OT, Hellhammer DH, Buske-Kirschbaum A, Von Auer K, Jobst S, et al. Free Cortisol Levels after Awakening: A Reliable Biological Marker for the Assessment of Adrenocortical Activity. *Life Sciences*. 1997;61(26):2539-49.
94. Jilma B, Hergovich N, Stohlawetz P, Eichler HG, Bauer P, Wagner OF. Circadian variation of granulocyte colony stimulating factor levels in man. *British Journal of Haematology*.106(2):368-70.
95. Birketvedt GS. Diurnal secretion of ghrelin, growth hormone, insulin binding proteins, and prolactin in normal weight and overweight subjects with and without the night eating syndrome. *Appetite*.59(3):688-93.
96. Cook CJ, Crewther BT. Within- and between-person variation in morning testosterone is associated with economic risk-related decisions in athletic women across the menstrual cycle. *Horm Behav*. 2019;112:77-80.
97. Cesari M, Penninx BW, Pahor M, Lauretani F, Corsi AM, Rhys Williams G, et al. Inflammatory markers and physical performance in older persons: the InCHIANTI study. *Journals of Gerontology Series A-Biological Sciences & Medical Sciences*. 2004;59(3):242-8.
98. Bull JR, Rowland SP, Scherwitzl EB, Scherwitzl R, Danielsson KG, Harper J. Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *NPJ Digit Med*. 2019;2:83.
99. Ghaffarilaleh G. Yoga positively affected depression and blood pressure in women with premenstrual syndrome in a randomized controlled clinical trial. *Complementary Therapies in Clinical Practice*.34:87-93.
100. Hansen L, Tjonneland A, Koster B, Brot C, Andersen R, Cohen AS, et al. Vitamin D Status and Seasonal Variation among Danish Children and Adults: A Descriptive Study. *Nutrients*. 2018;10(11).
101. Niculescu DA, Capatina CAM, Dusceac R, Caragheorgheopol A, Ghemigian A, Poiana C. Seasonal variation of serum vitamin D levels in Romania. *Arch Osteoporos*. 2017;12(1):113.
102. Robinson D, Bevan EA, Hinohara S, Takahashi T. Seasonal variation in serum cholesterol levels--evidence from the UK and Japan. *Atherosclerosis*. 1992;95(1):15-24.
103. Garde AH, Hansen AM, Skovgaard LT, Christensen JM. Seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulfate, hemoglobin A(1c), IgA, prolactin, and free testosterone in healthy women. *Clinical Chemistry*. 2000;46(4):551-9.
104. Westerterp KR. Seasonal variation in body weight: an experimental case study. *Journal of Thermal Biology*. 2001;26(4):525-7.
105. Modesti PA, Morabito M, Massetti L, Rapi S, Orlandini S, Mancina G, et al. Seasonal Blood Pressure Changes. *Hypertension*. 2013;61(4):908-14.
106. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biological variation: pros, cons and progress. *Scand J Clin Lab Invest*. 1999;59(7):491-500.
107. Ford RP. Essential data derived from biological variation for establishment and use of lipid analyses. *Annals of Clinical Biochemistry*. 1989;26(Pt 3):281-5.
108. Boey E, Gay GM, Poh KK, Yeo TC, Tan HC, Lee CH. Visit-to-visit variability in LDL- and HDL-cholesterol is associated with adverse events after ST-segment elevation myocardial infarction: A 5-year follow-up study. *Atherosclerosis*. 2016;244:86-92.
109. Yang D, Cai Q, Qi X, Zhou Y. Postprandial Lipid Concentrations and Daytime Biological Variation of Lipids in a Healthy Chinese Population. *Annals of Laboratory Medicine*. 2018;38(5):431-9.

110. Reed RG, Kris-Etherton P, Stewart PW, Pearson TA. Variation of lipids and lipoproteins in premenopausal women compared with men and postmenopausal women. DELTA (Dietary Effects on Lipoproteins and Thrombogenic Activity) Investigators. *Metabolism: Clinical & Experimental*. 2000;49(9):1101-5.
111. Idei M, Hirayama S, Miyake N, Kon M, Horiuchi Y, Ueno T, et al. Mean postprandial triglyceride concentration is an independent risk factor for carotid atherosclerosis in patients with type 2 diabetes. *Clinica Chimica Acta*. 2014;430:134-9.
112. Loh TP, Metz MP. Trends and physiology of common serum biochemistries in children aged 0-18 years. *Pathology*. 2015;47(5):452-61.
113. Liang L, He H, Zeng Y, Zhang M, Wang X, Li X, et al. Evaluation of biological variation of glycated hemoglobin and glycated albumin in healthy Chinese subjects. *Journal of Clinical Laboratory Analysis*. 2019;33(3):e22715.
114. Jorde R, Sundsfjord J. Intra-individual variability and longitudinal changes in glycaemic control in patients with Type 1 diabetes mellitus. *Diabetic Medicine*. 2000;17(6):451-6.
115. McCudden C, Akbari A, White CA, Biyani M, Hiremath S, Brown PA, et al. Individual patient variability with the application of the kidney failure risk equation in advanced chronic kidney disease. *PLoS ONE [Electronic Resource]*. 2018;13(6):e0198456.
116. Muntner P, Joyce C, Levitan EB, Holt E, Shimbo D, Webber LS, et al. Reproducibility of visit-to-visit variability of blood pressure measured as part of routine clinical care. *Journal of Hypertension*. 2011;29(12):2332-8.
117. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Inform Prim Care*. 2011;19(4):251-5.
118. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International Journal of Epidemiology*. 2019;48(6):1740-g.
119. Mahadevan P, Harley M, Fordyce S, Hodgson S, Ghosh R, Myles P, et al. Completeness and representativeness of small area socioeconomic data linked with the UK Clinical Practice Research Datalink (CPRD). *J Epidemiol Community Health*. 2022;76(10):880-6.
120. Shiekh SI, Harley M, Ghosh RE, Ashworth M, Myles P, Booth HP, et al. Completeness, agreement, and representativeness of ethnicity recording in the United Kingdom's Clinical Practice Research Datalink (CPRD) and linked Hospital Episode Statistics (HES). *Population Health Metrics*. 2023;21(1):3.
121. Persson R, Vasilakis-Scaramozza C, Hagberg KW, Sponholtz T, Williams T, Myles P, et al. CPRD Aurum database: Assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiology and Drug Safety*. 2020;29(11):1456-64.
122. Gokhale KM, Chandan JS, Toulis K, Gkoutos G, Tino P, Nirantharakumar K. Data extraction for epidemiological research (DExtER): a novel tool for automated clinical epidemiology studies. *Eur J Epidemiol*. 2021;36(2):165-78.
123. Barr RG, David MN, James BM, Daniel ES. Tests of Glycemia for the Diagnosis of Type 2 Diabetes Mellitus. *Ann Intern Med*. 2002;137(4):263.
124. Association AD. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021. *Diabetes Care*. 2021;44(Suppl 1):S15-s33.
125. Echouffo-Tcheugui JB, Selvin E. Prediabetes and What It Means: The Epidemiological Evidence. *Annual Review of Public Health*. 2021;42(Volume 42, 2021):59-77.
126. Anonymous. Type 2 diabetes in adults: management, NICE guideline [NG28] 2019 [Available from: <https://www.nice.org.uk/guidance/ng28>].
127. Lenters-Westra E, English E. Evaluating new HbA1c methods for adoption by the IFCC and NGSP reference networks using international quality targets. *Clin Chem Lab Med*. 2017;55(9):1426-34.

128. Dhath GS, Agarwal MM, Bishawi B. HbA1c: A comparison of NGSP with IFCC transformed values. *Clinica Chimica Acta*. 2005;358(1):81-6.
129. Lenters-Westra E, Roraas T, Schindhelm RK, Slingerland RJ, Sandberg S. Biological variation of Hemoglobin A1c: Consequences for diagnosing diabetes mellitus. *Nederlands Tijdschrift voor Klinische Chemie en Laboratoriumgeneeskunde*. 2015;40 (2):104.
130. Little RR, Rohlfing CL. The long and winding road to optimal HbA1c measurement. *Clin Chim Acta*. 2013;418:63-71.
131. Raphael A, Friger M, Biderman A. Seasonal variations in HbA1c among type 2 diabetes patients on a semi-arid climate between the years 2005-2015. *Prim Care Diabetes*. 2021;15(3):502-6.
132. Gonzalez-Lao E, Corte Z, Simon M, Ricos C, Coskun A, Braga F, et al. Systematic review of the biological variation data for diabetes related analytes. *Clinica Chimica Acta*. 2019;488:61-7.
133. Braga F, Dolci A, Mosca A, Panteghini M, Sacco L, editors. *Variabilità biologica delCE emoglobina glicata: un analisi sistematica della letteratura*2009.
134. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71-n.
135. Byron C. Wallace KS, Carla E. Brodley, Joseph Lau and Thomas A. Trikalinos, editor *Deploying an interactive machine learning system in an evidence-based practice center: abstrackr*. ACM International Health Informatics Symposium (IHI); 2012.
136. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol*. 2020;20(1):293.
137. Field AP. Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychol Methods*. 2005;10(4):444-67.
138. Wan EYF. Association of variability in hemoglobin A1c with cardiovascular diseases and mortality in Chinese patients with type 2 diabetes mellitus — A retrospective population-based cohort study. *Journal of Diabetes & its Complications*. 2016;30(7):1240-8.
139. Segar MW, Patel KV, Vaduganathan M, Caughey MC, Butler J, Fonarow GC, et al. Association of Long-term Change and Variability in Glycemia With Risk of Incident Heart Failure Among Patients With Type 2 Diabetes: A Secondary Analysis of the ACCORD Trial. *Diabetes care*. 2020;15.
140. Sugawara A, Kawai K, Motohashi S, Saito K, Kodama S, Yachi Y, et al. HbA1c variability and the development of microalbuminuria in type 2 diabetes: Tsukuba Kawai Diabetes Registry 2. *Diabetologia*. 2012;55(8):2128-31.
141. Mellergard E, Johnsson P, Eek F. Sociodemographic factors associated with HbA1c variability in type 2 diabetes: a prospective exploratory cohort study. *BMC Endocrine Disorders*. 2020;20(1).
142. Raman S, Dai H, DeLurgio SA, Williams DD, Lind M, Patton SR, et al. High hemoglobin A1c variability is associated with early risk of microalbuminuria in children with T1D. *Pediatr Diabetes*. 2016;17(6):398-406.
143. Yu ZB, Zhu Y, Li D, Wu MY, Tang ML, Wang JB, et al. Association between visit-to-visit variability of HbA_{1c} and cognitive decline: a pooled analysis of two prospective population-based cohorts. *Diabetologia*. 2020;63(1):85-94.
144. Carlsen S, Petersen PH, Skeie S, Oyvind S, Sandberg S. Within-subject biological variation of glucose and HbA1c in healthy persons and in type 1 diabetes patients. *Clinical Chemistry and Laboratory Medicine*. 2011;49(9):1501-7.
145. Maguire A, Mitchell BD, Ruzafa JC. Antihyperglycaemic treatment patterns, observed glycaemic control and determinants of treatment change among patients with type 2 diabetes

in the United Kingdom primary care: a retrospective cohort study. *BMC Endocr Disord*. 2014;14:73.

146. Hirst JA, Farmer AJ, Ali R, Roberts NW, Stevens RJ. Quantifying the effect of metformin treatment and dose on glycemic control. *Diabetes Care*. 2012;35(2):446-54.
147. Ma WY, Li HY, Pei D, Hsia TL, Lu KC, Tsai LY, et al. Variability in hemoglobin A1c predicts all-cause mortality in patients with type 2 diabetes. *Journal of Diabetes & its Complications*. 2012;26(4):296-300.
148. Caturano A, Galiero R, Pafundi PC, Cesaro A, Vetrano E, Palmiero G, et al. Does a strict glycemic control during acute coronary syndrome play a cardioprotective effect? Pathophysiology and clinical evidence. *Diabetes Res Clin Pract*. 2021;178:108959.
149. Chiu WC, Lai YR, Cheng BC, Huang CC, Chen JF, Lu CH. HbA1C Variability Is Strongly Associated with Development of Macroalbuminuria in Normal or Microalbuminuria in Patients with Type 2 Diabetes Mellitus: A Six-Year Follow-Up Study. *BioMed Research International*. 2020;2020 (no pagination).
150. Lui DTW, Lee CH, Chan YH, Chow WS, Fong CHY, Siu DCW, et al. HbA1c variability, in addition to mean HbA1c, predicts incident hip fractures in Chinese people with type 2 diabetes. *Osteoporosis International*. 2020.
151. Mo Y, Zhou J, Ma X, Zhu W, Zhang L, Li J, et al. Haemoglobin A1c variability as an independent correlate of atherosclerosis and cardiovascular disease in Chinese type 2 diabetes. *Diabetes & Vascular Disease Research*. 2018;15(5):402-8.
152. Cavagnoli G, Pimentel AL, Freitas PAC, Gross JL, Camargo JL. Effect of ethnicity on HbA1c levels in individuals without diabetes: Systematic review and meta-analysis. *PLoS One*. 2017;12(2):e0171315-e.
153. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15(2):155-63.
154. Aarsand AK, Roraas T, Fernandez-Calle P, Ricos C, Diaz-Garzon J, Jonker N, et al. The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation. *Clin Chem*. 2018;64(3):501-14.
155. Nagy-Simon T, Hada A-M, Suarasan S, Potara M. Recent advances on the development of plasmon-assisted biosensors for detection of C-reactive protein. *Journal of molecular structure*. 2021;1246:131178.
156. Ansar W, Ghosh S. C-reactive protein and the biology of disease. *Immunol Res*. 2013;56(1):131-42.
157. Casas JP, Shah T, Hingorani AD, Danesh J, Pepys MB. C-reactive protein and coronary heart disease: a critical review. *J Intern Med*. 2008;264(4):295-314.
158. Oppong R, Jit M, Smith RD, Butler CC, Melbye H, Mölstad S, et al. Cost-effectiveness of point-of-care C-reactive protein testing to inform antibiotic prescribing decisions. 2013.
159. Kaura A, Hartley A, Panoulas V, Glampson B, Shah ASV, Davies J, et al. Mortality risk prediction of high-sensitivity C-reactive protein in suspected acute coronary syndrome: A cohort study. *PLOS Medicine*. 2022;19(2):e1003911.
160. Nehring SM, Goyal A, Patel BC. C Reactive Protein. *StatPearls*. Treasure Island (FL): StatPearls Publishing

Copyright © 2022, StatPearls Publishing LLC.; 2022.

161. Pohanka M. Diagnoses Based on C-Reactive Protein Point-of-Care Tests. *Biosensors (Basel)*. 2022;12(5):344.
162. Eccles S, Pincus C, Higgins B, Woodhead M. Diagnosis and management of community and hospital acquired pneumonia in adults: summary of NICE guidance. *BMJ*. 2014;349(dec03 4):g6722-g.
163. Ghazizadeh H, Kathryn Bohn M, Ghaffarian Zirak R, Kamel Khodabandeh A, Zare-Feyzabadi R, Saberi-Karimian M, et al. Comprehensive laboratory reference intervals for routine

biochemical markers and pro-oxidant-antioxidant balance (PAB) in male adults. *J Clin Lab Anal*. 2020;34(11):e23470.

164. Cho SMJ, Lee H, Shim J-S, Kim HC. Sex-, Age-, and Metabolic Disorder-Dependent Distributions of Selected Inflammatory Biomarkers among Community-Dwelling Adults. *Diabetes & metabolism journal*. 2020;44(5):711-25.

165. Braga F, Panteghini M. Biologic variability of C-reactive protein: Is the available information reliable? *Clin Chim Acta*. 2012;413(15-16):1179-83.

166. Johnson TV, Abbasi A, Master VA. Systematic review of the evidence of a relationship between chronic psychosocial stress and C-reactive protein. *Mol Diagn Ther*. 2013;17(3):147-64.

167. Ain QU, Sarfraz M, Prasesti GK, Dewi TI, Kurniati NF. Confounders in Identification and Analysis of Inflammatory Biomarkers in Cardiovascular Diseases. *Biomolecules*. 2021;11(10).

168. Hilderink JM, van der Linden N, Kimenai DM, Litjens EJR, Klinkenberg LJJ, Aref BM, et al. Biological Variation of Creatinine, Cystatin C, and eGFR over 24 Hours. *Clin Chem*. 2018;64(5):851-60.

169. Barak-Corren Y, Barak-Corren N, Gileles-Hillel A, Heiman E. The effect of C-reactive protein on chest X-ray interpretation: A decision-making experiment among pediatricians. *Pediatr Pulmonol*. 2021;56(6):1644-50.

170. Povia P, Coelho L, Almeida E, Fernandes A, Mealha R, Moreira P, et al. Early identification of intensive care unit-acquired infections with daily monitoring of C-reactive protein: a prospective observational study. *Critical Care (London, England)*. 2006;10(2):R63.

171. Kelly K, Mezuk B. Depression, Inflammation, and Metabolic Risk: A Genetically-Informed Exploratory Study. *European Neuropsychopharmacology*. 2019;29 (Supplement 3):S976-S7.

172. Gibson GJ. Spirometry: then and now. *Breathe*. 2005;1(3):206-16.

173. Pierce R. Spirometry: an essential clinical measurement. *Australian Family Physician*. 2005;34(7):535-40.

174. Anonymous. Chronic obstructive pulmonary disease in adults. Quality Standard QS10 2016 [Available from: <https://www.nice.org.uk/guidance/qs10>.

175. Anonymous. Asthma: diagnosis, monitoring and chronic asthma management. NICE guideline [NG80] 2021 [Available from: <https://www.nice.org.uk/guidance/ng80/chapter/Recommendations#objective-tests-for-diagnosing-asthma-in-adults-young-people-and-children-aged-5-and-over>.

176. Barreiro TJ, Perillo I. An approach to interpreting spirometry. *Am Fam Physician*. 2004;69(5):1107-14.

177. Ekström M, Schiöler L, Grønseth R, Johannessen A, Svanes C, Leynaert B, et al. Absolute values of lung function explain the sex difference in breathlessness in the general population. *European Respiratory Journal*. 2017;49(5):1602047.

178. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J*. 2012;40(6):1324-43.

179. Yawn BP. Differential assessment and management of asthma vs chronic obstructive pulmonary disease. *Medscape J Med*. 2009;11(1):20.

180. Anonymous. National Institute for Health and Care Excellence Guideline. Diagnosis of chronic obstructive pulmonary disease 2023 [Available from: <https://cks.nice.org.uk/topics/chronic-obstructive-pulmonary-disease/diagnosis/diagnosis-copd/>.

181. Zhang X, Zhang Y, Zhou Y, Yin D, Lv C, Lin J, et al. Age-related circadian rhythm and variability of large- and small-airway function in healthy non-smoking adults: Data from 7-day diurnal and nocturnal home monitoring using an electronic portable spirometer. *Front Public Health*. 2022;10:946988.

182. McCormack MC, Shade D, Wise RA. Spirometer calibration checks : Is 3.5% good enough? *Chest*. 2007;131(5):1486-93.
183. Jensen RL, Teeter JG, England RD, White HJ, Pickering EH, Crapo RO. Instrument accuracy and reproducibility in measurements of pulmonary function. *Chest*. 2007;132(2):388-95.
184. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, et al. Standardisation of spirometry. *Eur Respir J*. 2005;26(2):319-38.
185. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, et al. Interpretative strategies for lung function tests. *Eur Respir J*. 2005;26(5):948-68.
186. Mohamed Hoesein FAA, Zanen P, Boezen HM, Groen HJM, van Ginneken B, de Jong PA, et al. Lung Function Decline in Male Heavy Smokers Relates to Baseline Airflow Obstruction Severity. *Chest*. 2012;142(6):1530-8.
187. Casanova C, Torres JPd, Aguirre-Jaime A, Pinto-Plata V, Marin JM, Cordoba E, et al. The Progression of Chronic Obstructive Pulmonary Disease Is Heterogeneous. *American Journal of Respiratory and Critical Care Medicine*. 2011;184(9):1015-21.
188. Kohansal R, Martinez-Camblor P, Agustí A, Buist AS, Mannino DM, Soriano JB. The Natural History of Chronic Airflow Obstruction Revisited. *American Journal of Respiratory and Critical Care Medicine*. 2009;180(1):3-10.
189. Woodruff PG, Barr RG, Bleecker E, Christenson SA, Couper D, Curtis JL, et al. Clinical Significance of Symptoms in Smokers with Preserved Pulmonary Function. *New England Journal of Medicine*. 2016;374(19):1811-21.
190. Ranganathan SC, Hoo AF, Lum SY, Goetz I, Castle RA, Stocks J. Exploring the relationship between forced maximal flow at functional residual capacity and parameters of forced expiration from raised lung volume in healthy infants. *Pediatr Pulmonol*. 2002;33(6):419-28.
191. Shin YH, Baek JH, Yoon JW, Jee HM, Yum HY, Choi SH, et al. Comparison of reference values and short-term variability for oscillatory and spirometric lung function in healthy korean preschool children. *Pediatric, Allergy, Immunology, and Pulmonology*. 2013;26(1):25-31.
192. Vilozni D, Barker M, Jellouschek H, Heimann G, Blau H. An interactive computer-animated system (SpiroGame) facilitates spirometry in preschool children. *American Journal of Respiratory & Critical Care Medicine*. 2001;164(12):2200-5.
193. Franca DC, Camargos PA, Martins JA, Abreu MC, Avelar e Araujo GH, Parreira VF. Feasibility and reproducibility of spirometry and inductance plethysmography in healthy Brazilian preschoolers. *Pediatr Pulmonol*. 2013;48(7):716-24.
194. Strachan DP. Repeatability of ventilatory function measurements in a population survey of 7 year old children. *Thorax*. 1989;44(6):474-9.
195. Fasola S, Ferrante G, Sabatini A, Santonico M, Zompanti A, Grasso S, et al. Repeatability of exhaled breath fingerprint collected by a modern sampling system in asthmatic and healthy children. *J Breath Res*. 2019;13(3):036007.
196. Minsky RC, Mucha FC, Keil PMR, Bobbio TG, Schivinski CIS. Reproducibility of spirometric maneuvers in healthy school children. *Pediatr Pulmonol*. 2016;51 (Supplement 42):S28.
197. Cooper PJ, Robertson CF, Hudson IL, Phelan PD. Variability of pulmonary function tests in cystic fibrosis. *Pediatr Pulmonol*. 1990;8(1):16-22.
198. Viegi G, Giuliano G, Mellocchi F, Pistelli G, Giuntini C. Variability of maximal expiratory flow-volume curve in young volley players. *Respiration*. 1988;54(1):33-41.
199. Scalco JC, Minsky RC, Schivinski CIS. Spirometry in Schoolchildren for Field Studies: Does Testing on Different Days Change the Result of the Exam? *Revista Paulista de Pediatria*. 2018;36(1):6.

200. Mallol J, Aguirre VL, Wandalsen G. Variability of the raised volume rapid thoracic compression technique in infants with recurrent wheezing. *Allergol Immunopathol (Madr)*. 2005;33(2):74-9.
201. Wu AC, Tantisira K, Li L, Schuemann B, Weiss S. Repeatability of response to asthma medications. *Journal of allergy and clinical immunology*. 2009;123(2):385-90.
202. Timonen KL, Randell JT, Salonen RO, Pekkanen J. Short-term variations in oscillatory and spirometric lung function indices among school children. *European Respiratory Journal*. 1997;10(1):82-7.
203. Kondo S, Ito M. Seasonal variation in the circadian rhythm of pulmonary function in stable asthmatic children who have nearly outgrown their asthma. *Chronobiology International*. 1996;13(4):295-303.
204. Jean-Baptiste JE, Amin N, Dumas ME, Dozor AJ. Spirometry performed with and without noseclips in children with asthma - Is there a difference? *Paediatric Respiratory Reviews*. 2010;1:S84.
205. Bastian-Lee Y, Chavasse R, Richter H, Seddon P. Assessment of a low-cost home monitoring spirometer for children. *Pediatr Pulmonol*. 2002;33(5):388-94.
206. Edmondson C, Westrupp N, Wallenburg J, Brownlee K, Alton EW, Bush A, et al. Monitoring lung function of young people with CF at home: Is it reliable? Results from the climb-CF study. *Pediatr Pulmonol*. 2020;55 (SUPPL 2):290.
207. Abbas C, Singer F, Yammine S, Casaulta C, Latzin P. Treatment response of airway clearance assessed by single-breath washout in children with cystic fibrosis. *Journal of Cystic Fibrosis*. 2013;12(6):567-74.
208. Sanders DB, Rosenfeld M, Mayer-Hamblett N, Stamey D, Redding GJ. Reproducibility of spirometry during cystic fibrosis pulmonary exacerbations. *Pediatr Pulmonol*. 2008;43(11):1142-6.
209. Iannaccone ST, Hynan LS. Reliability of 4 outcome measures in pediatric spinal muscular atrophy. *Arch Neurol*. 2003;60(8):1130-6.
210. Singer F, Schlegtendal A, Nyilas S, Latzin P, Boon M, Koerner-Rettberg C. Natural variability of lung clearance index and FEV1 in children and young adults with Primary Ciliary Dyskinesia. *European Respiratory Journal Conference: European Respiratory Society International Congress, ERS*. 2018;52(Supplement 62).
211. Connolly AM, Malkus EC, Mendell JR, Flanigan KM, Miller JP, Schierbecker JR, et al. Outcome reliability in non-ambulatory boys/men with Duchenne muscular dystrophy. *Muscle Nerve*. 2015;51(4):522-32.
212. Florence J, Connolly A, Philip Miller J, Malkus E, Schierbecker J, Siener C, et al. Outcomes measure reliability in non ambulatory boys and men with duchenne muscular dystrophy (DMD): Results from the muscular dystrophy association DMD clinical research network. *Neurology Conference: 64th American Academy of Neurology Annual Meeting New Orleans, LA United States Conference Publication*:. 2012;78(1 Meeting Abstract).
213. Hagmolen of ten Have W, van den Berg NJ, van der Palen J, Bindels PJ, van Aalderen WM. Validation of a single concentration methacholine inhalation provocation test (SCIPT) in children. *J Asthma*. 2005;42(6):419-23.
214. Nickerson BG, Lemen RJ, Gerdes CB, Wegmann MJ, Robertson G. Within-subject variability and per cent change for significance of spirometry in normal subjects and in patients with cystic fibrosis. *Am Rev Respir Dis*. 1980;122(6):859-66.
215. Kennedy MD, Steinback CD, Skow R, Parent EC. Is Performance of a Modified Eucapnic Voluntary Hyperpnea Test in High Ventilation Athletes Reproducible? *Allergy Asthma Immunol Res*. 2017;9(3):229-36.
216. Ross J, Bates DV, Dean E, Abboud RT. Discordance of airflow limitation and ventilatory inhomogeneity in asthma and cystic fibrosis. *Clinical & Investigative Medicine - Medecine Clinique et Experimentale*. 1992;15(2):97-102.

217. Franzinelli A, Angotzi G, Bellamio Sartorelli C. Reproducibility of maximum expiratory flow volume curve in normal subjects and in patients with asthmatic bronchitis. [Italian]. *Med Lav*. 1976;67(6):586-98.
218. Michoud MC, Ghezzi H, Amyot R. A comparison of pulmonary function tests used for bronchial challenges. *Clinical Respiratory Physiology*. 1982;18(4):609-21.
219. Williams NC, Johnson MA, Hunter KA, Sharpe GR. Reproducibility of the bronchoconstrictive response to eucapnic voluntary hyperpnoea. *Respir Med*. 2015;109(10):1262-7.
220. Cochrane GM, Prieto F, Clark TJ. Intrasubject variability of maximal expiratory flow volume curve. *Thorax*. 1977;32(2):171-6.
221. McCarthy DS, Craig DB, Cherniack RM. Intraindividual variability in maximal expiratory flow volume and closing volume in asymptomatic subjects. *Am Rev Respir Dis*. 1975;112(3):407-11.
222. Koh YI, In Seon C, Hyun Ju N, Seok Chae P, An Soo J. An evaluation of the accuracy of mini-Wright peak flow meter. [Korean]. *Tuberculosis and Respiratory Diseases*. 1997;44(2):298-308.
223. Randell JT, Salonen RO, Pekkarinen H, Tukiainen H. Short-term variations in oscillatory and spirometric lung function parameters of non-asthmatic adults. *Clinical Physiology*. 1999;19(4):329-37.
224. Gimeno F, van der Weele LT, Koeter GH, de Monchy JG, van Altena R. Variability of forced oscillation (Siemens Siregnost FD 5) measurements of total respiratory resistance in patients and healthy subjects. *Ann Allergy*. 1993;71(1):56-60.
225. Malmberg LP, Hedman J, Sovijarvi AR. Accuracy and repeatability of a pocket turbine spirometer: comparison with a rolling seal flow-volume spirometer. *Clinical Physiology*. 1993;13(1):89-98.
226. Zou K, Ni C, He Z, Zhang Y, Ma Z, Song J, et al. [Reproducibility of a portable spirometer based on differential pressure sensor]. *Zhongguo Yi Liao Qi Xie Za Zhi*. 2013;37(1):30-2, 48.
227. Chinn DJ, Cotes JE. Transit time indices derived from forced expiratory spiograms: repeatability and criteria for curve selection and truncation. *European Respiratory Journal*. 1994;7(2):402-8.
228. Lebowitz MD, Knudson RJ, Robertson G, Burrows B. Significance of intraindividual changes in maximum expiratory flow volume and peak expiratory flow measurements. *Chest*. 1982;81(5):566-70.
229. Lam S, Abboud RT, Chan-Yeung M, Tan F. Use of maximal expiratory flow-volume curves with air and helium-oxygen in the detection of ventilatory abnormalities in population surveys. *Am Rev Respir Dis*. 1981;123(2):234-7.
230. Lawther PJ, Brooks AG, Lord PW, Waller RE. Day-to-day changes in ventilatory function in relation to the environment. Part III. Frequent measurements of peak flow. *Environmental Research*. 1974;8(1):119-30.
231. Struthers AD, Addis GJ. Respiratory function measurements in clinical pharmacological studies including an assessment of the area under the MEFV curve as a new parameter in chronic bronchitic patients. *European Journal of Clinical Pharmacology*. 1988;34(3):277-81.
232. Chinn DJ, Lee WR. Within- and between- subject variability of indices from the closing volume and flow volume traces. *Bulletin European de Physiopathologie Respiratoire*. 1977;13(6):789-802.
233. Neild JE, Twort CH, Chinn S, McCormack S, Jones TD, Burney PG, et al. The repeatability and validity of respiratory resistance measured by the forced oscillation technique. *Respir Med*. 1989;83(2):111-8.
234. Humerfelt S, Eide GE, Kvale G, Gulsvik A. Forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC) variability in asymptomatic never-smoking men. *Clinical Physiology*. 1998;18(4):387-96.

235. Rozas CJ, Goldman AL. Daily spirometric variability: normal subjects and subjects with chronic bronchitis with and without airflow obstruction. *Archives of Internal Medicine*. 1982;142(7):1287-91.
236. Wang ML, Gunel E, Petsonk EL. Design strategies for longitudinal spirometry studies: study duration and measurement frequency. *American Journal of Respiratory & Critical Care Medicine*. 2000;162(6):2134-8.
237. Gulsvik A, Tosteson T, Bakke P, Humerfelt S, Weiss ST, Speizer FE. Expiratory and inspiratory forced vital capacity and one-second forced volume in asymptomatic never-smokers in Norway. *Clinical Physiology*. 2001;21(6):648-60.
238. Jensen RL, Teeter JG, England RD, Howell HM, White HJ, Pickering EH, et al. Sources of long-term variability in measurements of lung function: implications for interpretation and clinical trial design. *Chest*. 2007;132(2):396-402.
239. Baur X, Degens P, Heitmann R, Hillenbach C, Marek W, Rausch V, et al. Lung function testing: The dilemma of predicted values in relation to the individual variability. *Respiration*. 1996;63(3):123-30.
240. Kainu A, Lindqvist A, Sarna S, Sovijarvi A. Intra-session repeatability of FET and FEV6 in the general population. *Clin Physiol Funct Imaging*. 2008;28(3):196-201.
241. Enright PL, Johnson LR, Connett JE, Voelker H, Buist AS. Spirometry in the Lung Health Study. 1. Methods and quality control. *Am Rev Respir Dis*. 1991;143(6):1215-23.
242. Hamedani H, Kadlecsek SJ, Ishii M, Emami K, Kuzma NN, Xin Y, et al. A variability study of regional alveolar oxygen tension measurement in humans using hyperpolarized ³He MRI. *Magnetic Resonance in Medicine*. 2013;70(6):1557-66.
243. Beckett WS, Pace PA, Sferlazza SJ, Carey VJ, Weiss ST. Annual variability in methacholine responsiveness in nonasthmatic working adults. *European Respiratory Journal*. 1997;10(11):2515-21.
244. Biddiscombe MF, Verbanck S, Meah S, Barnes PJ, Usmani OS. Reproducibility of the multiple breath nitrogen washout test in healthy volunteers and subjects with mild to moderate asthma. *Thorax*. 2009;64:A128.
245. Billings C, Fishwick D. Spirometry longitudinal data analysis software (SPIROLA) in a routine clinical laboratory. *European Respiratory Journal Conference: European Respiratory Society Annual Congress*. 2014;44(SUPPL. 58).
246. Burki NK, Dent MC. The forced expiratory time as a measure of small airway resistance. *Clin Sci Mol Med*. 1976;51(1):53-8.
247. Garrard CS, Emmons C. The reproducibility of the respiratory responses to maximum exercise. *Respiration*. 1986;49(2):94-100.
248. Houghton CM. Lung function measurements in asthma clinical trials: The University of Manchester (United Kingdom); 2005.
249. Liu X, Jurisevic M, Keatley D, Kidd P, Smith B. A comparison of medgraphics and Jaeger lung function equipment with spirometry, diffusing capacity and bodyplethysmography parameters by lung function biocontrol measurements. *Respirology*. 2018;23 (Supplement 1):13.
250. Wise RA, Connett J, Kurnow K, Grill J, Johnson L, Kanner R, et al. Selection of spirometric measurements in a clinical trial, the Lung Health Study. *American Journal of Respiratory and Critical Care Medicine*. 1995;151(3 I):675-81.
251. Rowan SA, Bradley JM, Bradbury I, Lawson J, Lynch T, Gustafsson P, et al. Lung clearance index is a repeatable and sensitive indicator of radiological changes in bronchiectasis. *American Journal of Respiratory & Critical Care Medicine*. 2014;189(5):586-92.
252. Holmner A, Ohberg F, Wiklund U, Bergmann E, Blomberg A, Wadell K. How stable is lung function in patients with stable chronic obstructive pulmonary disease when monitored using a telehealth system? A longitudinal and home-based study. *BMC Med Inf Decis Mak*. 2020;20(1):87.

253. Taube C, Kannie F, Gronke L, Richter K, Mucke M, Paasch K, et al. Reproducibility of forced inspiratory and expiratory volumes after bronchodilation in patients with COPD or asthma. *Respir Med*. 2003;97(5):568-77.
254. Owens MW, Kinasewitz GT, Strain DS. Evaluating the effects of chronic therapy in patients with irreversible air-flow obstruction. *Am Rev Respir Dis*. 1986;134(5):935-7.
255. Anderson WH, Ha JW, Couper DJ, O'Neal WK, Barr RG, Bleecker ER, et al. Variability in objective and subjective measures affects baseline values in studies of patients with COPD. *PLoS ONE*. 2017;12(9).
256. Mahler DA, Ward J, Waterman LA, McCusker C, ZuWallack R, Baird JC. Patient-reported dyspnea in COPD reliability and association with stage of disease. *Chest*. 2009;136(6):1473-9.
257. O'Donnell DE, Lam M, Webb KA. Measurement of symptoms, lung hyperinflation, and endurance during exercise in chronic obstructive pulmonary disease. *American Journal of Respiratory & Critical Care Medicine*. 1998;158(5 Pt 1):1557-65.
258. Nosedá A, Carpioux JP, Prigogine T, Schmerber J. Lung function, maximum and submaximum exercise testing in COPD patients: reproducibility over a long interval. *Lung*. 1989;167(4):247-57.
259. Bofan M, Mores N, Baron M, Dabrowska M, Valente S, Schmid M, et al. Within-day and between-day repeatability of measurements with an electronic nose in patients with COPD. *J Breath Res*. 2013;7(1):017103.
260. Sirichana W, Patel MH, Wang X, Taylor M, Barjaktarevic I, Kleerup EC, et al. Choices of spirometry measures for remote patient monitoring in COPD. *American Journal of Respiratory and Critical Care Medicine Conference: American Thoracic Society International Conference, ATS*. 2014;189(MeetingAbstracts).
261. Chan-Thim E, Dumont M, Moullec G, Rizk AK, Wardini R, Trutschnigg B, et al. Clinical impact of time of day on acute exercise response in COPD. *Copd*. 2014;11(2):204-11.
262. Laube BL, Carson KA, Sharpless G, Paulin LM, Hansel NN. Mucociliary Clearance in Former Tobacco Smokers with Both Chronic Obstructive Pulmonary Disease and Chronic Bronchitis and the Effect of Roflumilast. *J Aerosol Med Pulm Drug Deliv*. 2019;32(4):189-99.
263. Yamamoto M, Konno S, Makita H, Nitta K, Shimizu K, Suzuki M, et al. Factors for the variability of three acceptable maximal expiratory flow-volume curves in chronic obstructive pulmonary disease. *International Journal of COPD*. 2021;16:415-22.
264. van Noord JA, Clement J, van de Woestijne KP, Demedts M. Total respiratory resistance and reactance as a measurement of response to bronchial challenge with histamine. *Am Rev Respir Dis*. 1989;139(4):921-6.
265. Faul JL, Demers EA, Burke CM, Poulter LW. The reproducibility of repeat measures of airway inflammation in stable atopic asthma. *American Journal of Respiratory & Critical Care Medicine*. 1999;160(5 Pt 1):1457-61.
266. Huang C, Izmailova ES, Jackson N, Ellis R, Bhatia G, Ruddy M, et al. Remote FEV1 Monitoring in Asthma Patients: A Pilot Study. *Clinical and Translational Science*. 2021;14(2):529-35.
267. Shrestha Palikhe N, Bosonea AM, Laratta C, Gandhi VD, Nahirney D, Hillaby A, et al. Stability of peripheral blood immune markers in patients with asthma. *Allergy Asthma Clin Immunol*. 2019;15:30.
268. Jung M, Korn S, Taube C, Buhl R. Short-term reproducibility of non-invasive clinical and inflammatory parameters in asthma. *American Journal of Respiratory and Critical Care Medicine Conference: American Thoracic Society International Conference, ATS*. 2011;183(1 MeetingAbstracts).
269. Roberto DM, Scarlett GA, Ana MI, Laura FT, Pablo BV, Solange CL. Spirometry interpretation feasibility among pre-school children according to the European respiratory society and american thoracic society guidelines. [Spanish]. *Rev Chil Pediatr*. 2015;86(2):86-91.

270. Donaire RM, Gonzalez SA, Moya AI, Fierro LT, Brockmann PV, Caussade SL. Spirometry interpretation feasibility among pre-school children according to the European Respiratory Society and American Thoracic Society Guidelines. *Rev Chil Pediatr.* 2015;86(2):86-91.
271. Alamer MA, Bhatawadekar SA, Watson WT, Maksym GN. Repeatability of respiratory impedance and bronchodilatory response in asthmatic children. *American Journal of Respiratory and Critical Care Medicine Conference: American Thoracic Society International Conference, ATS.* 2015;191(MeetingAbstracts).
272. Baptist AP, Sengupta R, Pranathiageswaran S, Wang Y, Ager J. Evaluation of exhaled nitric oxide measurements in the emergency department for patients with acute asthma. *Ann Allergy Asthma Immunol.* 2008;100(5):415-9.
273. Barley EA, Jones PW. Repeatability of a Rasch model of the AQ20 over five assessments. *Qual Life Res.* 2006;15(5):801-9.
274. Dennis KA, Moore V, Burge C, Burge S. Is FEV1 better than PEF at detecting asthmatic changes in diurnal variation? *European Respiratory Journal Conference: European Respiratory Society Annual Congress.* 2011;38(SUPPL. 55).
275. Silkoff P, Laviolette M, Singh D, FitzGerald J, Kelsen S, Backer V, et al. Longitudinal stability of asthma characteristics and biomarkers from the Airways Disease Endotyping for Personalized Therapeutics (ADEPT) study. *Respiratory research.* 2016;17(1):43.
276. Stanbrook MB, Corey M, Tullis DE. The repeatability of forced expiratory volume measurements in adults with cystic fibrosis. *Chest.* 2004;125(1):150-5.
277. White H, Shaw N, Denman S, Pollard K, Wynne S, Peckham DG. Variation in lung function as a marker of adherence to oral and inhaled medication in cystic fibrosis. *European Respiratory Journal.* 2017;49(3):03.
278. White H, Shaw N, Gillgrass L, Wood A, Chadwick HK, Peckham D. Evaluation of an RCT web based intervention for adherence in cystic fibrosis. *Pediatr Pulmonol.* 2017;52 (Supplement 47):503.
279. Bayfield KJ, Alton EW, Boyd A, Cheng SH, Cunningham S, Davies JC, et al. Lung function outcome measures in a multidose trial of non-viral gene therapy in patients with CF. *Pediatr Pulmonol.* 2014;49:364-5.
280. Shaw N, White H, Denman S, Peckham D. The impact of disease severity and clinical variation on self-reported adherence. *Journal of Cystic Fibrosis.* 2014;13:S117.
281. Lindgren BR, Finkelstein SM, Prasad B, Dutta P, Killoren T, Scherber J, et al. Determination of reliability and validity in home monitoring data of pulmonary function tests following lung transplantation. *Res Nurs Health.* 1997;20(6):539-50.
282. Kafaja S, Clements PJ, Wilhalme H, Tseng CH, Furst DE, Kim GH, et al. Reliability and minimal clinically important differences of FVC results from the scleroderma lung studies (SLS-I and SLS-II). *American journal of respiratory and critical care medicine.* 2018;197(5):644-52.
283. Sheshadri A, Balagani A, Alousi AM, Bashoura L, Stolar K, Arain MH, et al. Feasibility and validity of a home-based spirometry program after hematopoietic stem cell transplantation for the early detection of pulmonary impairment. *American Journal of Respiratory and Critical Care Medicine Conference: American Thoracic Society International Conference, ATS.* 2018;197(MeetingAbstracts).
284. Moor CC, Van Leuven SI, Wijsenbeek MS, Vonk MC. Feasibility of online home spirometry in systemic sclerosis-associated interstitial lung disease: A pilot study. *Rheumatology (United Kingdom).* 2021;60(5):2467-71.
285. Tan GP, Soon LHY, Ni B, Cheng H, Tan AKH, Kor AC, et al. Variability of forced expiratory manoeuvre in patients with advanced neuromuscular diseases. *Respirology.* 2017;22 (Supplement 3):151-2.
286. Krug LM, Arduino JM, Sun X, Kindler HL, Manegold C, Fennell D, et al. Forced vital capacity (FVC) as a reproducible measure of pulmonary function (PF) in chemotherapy-

- pretreated patients with malignant pleural mesothelioma (MPM). *Journal of Clinical Oncology Conference: ASCO Annual Meeting*. 2011;29(15 SUPPL. 1).
287. Berry G. Longitudinal observations. Their usefulness and limitations with special reference to the forced expiratory volume. *BullPhysio-PathResp*. 1974;10(5):643-56.
 288. Glindmeyer HW, Anderson ST, Diem JE, Weill H. A comparison of the Jones and Stead-Wells spirometers. *Chest*. 1978;73(5):596-62.
 289. Fonseca JA, Costa-Pereira A, Delgado L, Silva LN, Magalhaes M, Castel-Branco MG, et al. Pulmonary function electronic monitoring devices: a randomized agreement study. *Chest*. 2005;128(3):1258-65.
 290. Moore VC, Parsons NR, Jaakkola MS, Burge CB, Pantin CF, Robertson AS, et al. Serial lung function variability using four portable logging meters. *J Asthma*. 2009;46(9):961-6.
 291. Klarholt E, Andersen LH, Thomsen A. A new, simple apparatus for the measurement of ventilatory function (Airflometer). [Danish]. *Ugeskrift for Laeger*. 1979;141(13):844-6.
 292. De Vries R, Dagelet JWF, Frey U, Lutter R, Maitland-Van Der Zee AH, Sterk PJ, et al. Assessment of repeatability of eNose (SpiroNose) measurements in healthy and asthmatic subjects. *European Respiratory Journal Conference: European Respiratory Society International Congress, ERS*. 2018;52(Supplement 62).
 293. Rossman MD, Xu Y, Kadlecsek S, Xin Y, Emami K, Ishii M, et al. Reproducibility assessment of hyperpolarized gas diffusion MRI compared to pulmonary function tests for diagnostic applications. *American Journal of Respiratory and Critical Care Medicine Conference: American Thoracic Society International Conference, ATS*. 2011;183(1 MeetingAbstracts).
 294. Yawn BP. Factors accounting for asthma variability: achieving optimal symptom control for individual patients. *Prim*. 2008;17(3):138-47.
 295. Cazzola M, MacNee W, Martinez FJ, Rabe KF, Franciosi LG, Barnes PJ, et al. Outcomes for COPD pharmacological trials: from lung function to biomarkers. *Eur Respir J*. 2008;31(2):416-69.
 296. Anonymous. *Depression, the Treatment and Management of Depression in Adults (Updated edition)*. Leicester: The British Psychological Society and The Royal College of Psychiatrists; 2019.
 297. Lakkis NA. Screening instruments for depression in primary care: a concise review for clinicians. *Postgraduate Medicine*. 127(1):99-107.
 298. American Psychiatric Association aQAPA. *Diagnostic and statistical manual of mental disorders : DSM-5 / American Psychiatric Association*. Fifth edition. ed: Washington, District of Columbia
London, England : American Psychiatric Publishing, 2013.; 2013.
 299. Osório FL, Loureiro SR, Hallak JEC, Machado-de-Sousa JP, Ushirohira JM, Baes CVW, et al. Clinical validity and intrarater and test-retest reliability of the Structured Clinical Interview for DSM-5 – Clinician Version (SCID-5-CV). *Psychiatry Clin Neurosci*. 2019;73(12):754-60.
 300. Brodey B, Purcell SE, Rhea K, Maier P, First M, Zweede L, et al. Rapid and Accurate Behavioral Health Diagnostic Screening: Initial Validation Study of a Web-Based, Self-Report Tool (the SAGE-SR). *J Med Internet Res*. 2018;20(3):e108.
 301. British Medical Association Ne. *Revisions to the GMS contract, 2006/7: delivering investment in general practice*. London: BMA; 2006.
 302. Anonymous. *NICE clinical guideline CG90: Depression in adults: recognition and management 2009* [Available from: <https://www.nice.org.uk/guidance/cg90>].
 303. Matheson GJ, Schain M, Almeida R, Lundberg J, Cselényi Z, Borg J, et al. Diurnal and seasonal variation of the brain serotonin system in healthy male subjects. *Neuroimage*. 2015;112:225-31.
 304. Sitch AJ. MODELLING OPTIMAL USE OF TESTS FOR MONITORING

DISEASE PROGRESSION AND RECURRENCE: University of Birmingham; 2019.

305. Kendrick T, Dowrick C, McBride A, Howe A, Clarke P, Maisey S, et al. Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ*. 2009;338(mar19 1):b750-b.
306. Karmisholt J, Andersen S. Detecting True Change in the Hospital Anxiety and Depression Scale, SF-36, and Hypothyroid Score when Monitoring Patients with Subclinical Hypothyroidism. *European Thyroid Journal*. 2019;8(3):144-51.
307. Shelton AT, Houghton NY, Morris DO, Latchford GL, Bekker HL, Munyombwe T. The development and validation of a psychological questionnaire for patients undergoing orthognathic treatment. *Orthodontics & craniofacial research*. 2015;18(1):51-64.
308. Weobong B, Akpalu B, Doku V, Owusu-Agyei S, Hurt L, Kirkwood B, et al. The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana. *Journal of Affective Disorders*. 2009;113(1-2):109-17.
309. Löwe B. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical Care*. 2004;42(12):1194-202.
310. Duan-Porter W, Hatch D, Pendergast JF, Freude G, Rose U, Burr H, et al. 12-month trajectories of depressive symptoms among nurses-Contribution of personality, job characteristics, coping, and burnout. *Journal of Affective Disorders*. 2018;234:67-73.
311. Beswick E, Quigley S, Macdonald P, Patrick S, Colville S, Chandran S, et al. The Patient Health Questionnaire (PHQ-9) as a tool to screen for depression in people with multiple sclerosis: a cross-sectional validation study. *BMC psychology*. 2022;10(1):281.
312. Delgadillo J, Payne S, Gilbody S, Godfrey C, Gore S, Jessop D, et al. How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. *Journal of Affective Disorders*. 2011;134(1-3):266-71.
313. Hyland KA, Hoogland AI, Gonzalez BD, Nelson AM, Lechner S, Tyson DM, et al. Evaluation of the Psychometric and Structural Properties of the Spanish Version of the Hospital Anxiety and Depression Scale in Latina Cancer Patients. *Journal of Pain and Symptom Management*. 2019;58(2):289-96.e2.
314. Quek KF, Low WY, Razack AH, Loh CS. Beck Depression Inventory (BDI): a reliability and validity test in the Malaysian urological population. *The Medical journal of Malaysia*. 2001;56(3):285-92.
315. Yusoff N, Low WY, Yip CH. Psychometric properties of the Malay Version of the Hospital Anxiety and Depression Scale: A study of husbands of breast cancer patients in Kuala Lumpur, Malaysia. *Asian Pacific Journal of Cancer Prevention*. 2011;12(4):915-7.
316. Silverberg J, Lee B, Lei D, Yousaf M, Janmohamed S, Mann C, et al. PMH39 Measurement Properties of Patient Health Questionnaire (PHQ)-9 and PHQ-2 in Adult Patients with Atopic Dermatitis. *Value Health*. 2020;23(Supplement 2):S590-S1.
317. Tiksnadi BB, Triani N, Fihaya FY, Turu' Allo IJ, Iskandar S, Putri DAE. Validation of Hospital Anxiety and Depression Scale in an Indonesian population: a scale adaptation study. *Family medicine and community health*. 2023;11(2).
318. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42.
319. Dong W, Zhang S, Yan S, Zhao Z, Zhang Z, Gu W. Clinical characteristics of patients with early-onset diabetes mellitus: a single-center retrospective study. *BMC Endocr Disord*. 2023;23(1):216.
320. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care*. 2020;44(Supplement_1):S15-S33.
321. Federation ID. IDF Diabetes Atlas, 10th edn. Brussels,Belgium: International Diabetes Federation, 2021. Online version International Diabetes Federation.IDF Diabetes Atlas, 10th edn. Brussels,Belgium: 2021. 2021.

322. Rooney MR, Fang M, Ogurtsova K, Ozkan B, Echouffo-Tcheugui JB, Boyko EJ, et al. Global Prevalence of Prediabetes. *Diabetes Care*. 2023;46(7):1388-94.
323. Nathan DM, Turgeon H, Regan S. Relationship between glycated haemoglobin levels and mean glucose levels over time. *Diabetologia*. 2007;50(11):2239-44.
324. Weykamp C, John WG, Mosca A, Hoshino T, Little R, Jeppsson JO, et al. The IFCC Reference Measurement System for HbA1c: a 6-year progress report. *Clin Chem*. 2008;54(2):240-8.
325. Bansal N. Prediabetes diagnosis and treatment: A review. *World J Diabetes*. 2015;6(2):296-303.
326. Anonymous. European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Biological Variation Database 2022 [Available from: <https://biologicalvariation.eu/about>].
327. Gough A, Sitch A, Ferris E, Marshall T. Within-subject variation of HbA1c: A systematic review and meta-analysis. *PLoS One*. 2023;18(8):e0289085.
328. National Institute for Clinical Excellence. When should I suspect type 2 diabetes in an adult? 2023 [Available from: <https://cks.nice.org.uk/topics/diabetes-type-2/diagnosis/diagnosis-in-adults/>].
329. Pivovarov R, Albers DJ, Hripcsak G, Sepulveda JL, Elhadad N. Temporal trends of hemoglobin A1c testing. *Journal of the American Medical Informatics Association*. 2014;21(6):1038-44.
330. Sidorenkov G, Haaijer-Ruskamp FM, de Zeeuw D, Denig P. A Longitudinal Study Examining Adherence to Guidelines in Diabetes Care According to Different Definitions of Adequacy and Timeliness. *PLOS ONE*. 2011;6(9):e24278.
331. Carson AP, Fox CS, McGuire DK, Levitan EB, Laclaustra M, Mann DM, et al. Low hemoglobin A1c and risk of all-cause mortality among US adults without diabetes. *Circ Cardiovasc Qual Outcomes*. 2010;3(6):661-7.
332. Booth HP, Prevost AT, Gulliford MC. Epidemiology of clinical body mass index recording in an obese population in primary care: a cohort study. *J Public Health (Oxf)*. 2013;35(1):67-74.
333. Lenters-Westra E, Schindhelm RK, Bilo HJ, Groenier KH, Slingerland RJ. Differences in interpretation of haemoglobin A1c values among diabetes care professionals. *Neth J Med*. 2014;72(9):462-6.
334. Jarman B, Townsend P, Carstairs V. Deprivation indices. *Bmj*. 1991;303(6801):523.
335. Santos C, Dias C. Note on the coefficient of variation properties. *Brazilian Electronic Journal of Mathematics*. 2021;2(4):101-11.
336. Durlach V, Vergès B, Al-Salameh A, Bahougne T, Benzerouk F, Berlin I, et al. Smoking and diabetes interplay: A comprehensive review and joint statement. *Diabetes Metab*. 2022;48(6):101370.
337. Walker RJ, Strom Williams J, Egede LE. Influence of Race, Ethnicity and Social Determinants of Health on Diabetes Outcomes. *Am J Med Sci*. 2016;351(4):366-73.
338. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479.
339. National INstitute for Clinical Excellence. Type 2 diabetes in adults: management, NICE guideline [NG28] 2022 [Available from: <https://www.nice.org.uk/guidance/ng28/chapter/Recommendations>].
340. Niederman MS. Recent advances in community-acquired pneumonia: inpatient and outpatient. *Chest*. 2007;131(4):1205-15.
341. Lei R, Mohan C. Immunological Biomarkers of COVID-19. *Crit Rev Immunol*. 2020;40(6):497-512.
342. Abdelhafiz D, Baker T, Glasgow DA, Abdelhafiz A. Biomarkers for the diagnosis and treatment of rheumatoid arthritis - a systematic review. *Postgrad Med*. 2023;135(3):214-23.

343. Ma J, Li L, Qie X, Zhao Q, Zhang L, Xu N, et al. Value of Combined Detection of PCT, CRP, and FIB in Differentiating Viral Infection from Bacterial Infection in Severe Pneumonia. *Clin Lab*. 2023;69(11).
344. Pepys MB, Hirschfield GM. C-reactive protein: a critical update. *The Journal of Clinical Investigation*. 2003;111(12):1805-12.
345. Denegri A, Boriani G. High Sensitivity C-reactive Protein (hsCRP) and its Implications in Cardiovascular Outcomes. *Curr Pharm Des*. 2021;27(2):263-75.
346. Ananthakrishnan AN, Adler J, Chachu KA, Nguyen NH, Siddique SM, Weiss JM, et al. AGA Clinical Practice Guideline on the Role of Biomarkers for the Management of Crohn's Disease. *Gastroenterology*. 2023;165(6):1367-99.
347. Landry A, Docherty P, Ouellette S, Cartier LJ. Causes and outcomes of markedly elevated C-reactive protein levels. *Can Fam Physician*. 2017;63(6):e316-e23.
348. Canchola J, Tang S, Hemyari P, Paxinos E, Marins E. Correct use of percent coefficient of variation (% CV) formula for log-transformed data. *MOJ Proteomics Bioinform*. 2017;6(4):316-7.
349. Petel D, Winters N, Gore GC, Papenburg J, Beltempo M, Lacroix J, et al. Use of C-reactive protein to tailor antibiotic use: a systematic review and meta-analysis. *BMJ Open*. 2018;8(12):e022133.
350. Hensor EM, Emery P, Bingham SJ, Conaghan PG. Discrepancies in categorizing rheumatoid arthritis patients by DAS-28(ESR) and DAS-28(CRP): can they be reduced? *Rheumatology (Oxford)*. 2010;49(8):1521-9.
351. Levy ML, Quanjer PH, Booker R, Cooper BG, Holmes S, Small I. Diagnostic spirometry in primary care: Proposed standards for general practice compliant with American Thoracic Society and European Respiratory Society recommendations: a General Practice Airways Group (GPIAG)1 document, in association with the Association for Respiratory Technology & Physiology (ARTP)2 and Education for Health3 1 www.gpiag.org 2 www.artp.org 3 www.educationforhealth.org.uk. *Prim Care Respir J*. 2009;18(3):130-47.
352. van de Hei SJ, Flokstra-de Blok BMJ, Baretta HJ, Doornewaard NE, van der Molen T, Patberg KW, et al. Quality of spirometry and related diagnosis in primary care with a focus on clinical use. *NPJ Prim Care Respir Med*. 2020;30(1):22.
353. Gore C, Crockett A, Pederson D, Booth M, Bauman A, Owen N. Spirometric standards for healthy adult lifetime nonsmokers in Australia. *European Respiratory Journal*. 1995;8(5):773-82.
354. (NICE) NifHaCE. Chronic obstructive pulmonary disease in over 16s: diagnosis and management 2019 [Available from: <https://www.nice.org.uk/guidance/ng115/chapter/Recommendations#diagnosing-copd>.
355. Koegelenberg CFN, Swart F, Irusen EM. Guideline for office spirometry in adults, 2012 : guideline. *South African Medical Journal*. 2013;103(1):52-61.
356. Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, et al. Standardization of Spirometry 2019 Update. An Official American Thoracic Society and European Respiratory Society Technical Statement. *Am J Respir Crit Care Med*. 2019;200(8):e70-e88.
357. Tashkin DP. Variations in FEV₁ decline over time in chronic obstructive pulmonary disease and its implications. *Curr Opin Pulm Med*. 2013;19(2):116-24.
358. Mazic S, Lazovic B, Djelic M, Suzic-Lazic J, Djordjevic-Saranovic S, Durmic T, et al. Respiratory parameters in elite athletes – does sport have an influence? *Pulmonology*. 2015;21(4):192-7.
359. Al Ghobain M. The effect of obesity on spirometry tests among healthy non-smoking adults. *BMC Pulm Med*. 2012;12:10.
360. Wang S, Sun X, Hsia TC, Lin X, Li M. The effects of body mass index on spirometry tests among adults in Xi'an, China. *Medicine (Baltimore)*. 2017;96(15):e6596.

361. Donohue JF. Minimal Clinically Important Differences in COPD Lung Function. *Copd*. 2005;2(1):111-24.
362. Rea H, Kenealy T, Adair J, Robinson E, Sheridan N. Spirometry for patients in hospital and one month after admission with an acute exacerbation of COPD. *International Journal of Chronic Obstructive Pulmonary Disease*. 2011;6(null):527-32.
363. Yawn BP, Enright PL, Lemanske RF, Israel E, Pace W, Wollan P, et al. Spirometry Can Be Done in Family Physicians' Offices and Alters Clinical Decisions in Management of Asthma and COPD. *Chest*. 2007;132(4):1162-8.
364. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G, et al. Cross-national epidemiology of DSM-IV major depressive episode. *BMC Medicine*. 2011;9(1):90.
365. Malhi GS, Mann JJ. Depression. *Lancet*. 2018;392(10161):2299-312.
366. Malhi GS, Outhred T, Hamilton A, Boyce PM, Bryant R, Fitzgerald PB, et al. Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for mood disorders: major depression summary. *Med J Aust*. 2018;208(4):175-80.
367. (NICE) NIfHaCE. Depression in adults: treatment and management 2022 [Available from: <https://www.nice.org.uk/guidance/ng222>.
368. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606-13.
369. Hudgens S, Floden L, Blackowicz M, Jamieson C, Popova V, Fedgchin M, et al. Meaningful Change in Depression Symptoms Assessed with the Patient Health Questionnaire (PHQ-9) and Montgomery-Åsberg Depression Rating Scale (MADRS) Among Patients with Treatment Resistant Depression in Two, Randomized, Double-blind, Active-controlled Trials of Esketamine Nasal Spray Combined With a New Oral Antidepressant. *J Affect Disord*. 2021;281:767-75.
370. Turkoz I, Alphs L, Singh J, Jamieson C, Daly E, Shawi M, et al. Clinically meaningful changes on depressive symptom measures and patient-reported outcomes in patients with treatment-resistant depression. *Acta Psychiatr Scand*. 2021;143(3):253-63.
371. Button KS, Kounali D, Thomas L, Wiles NJ, Peters TJ, Welton NJ, et al. Minimal clinically important difference on the Beck Depression Inventory - II according to the patient's perspective. *Psychological Medicine*. 2015;45(15):3269-79.
372. Kounali D, Button KS, Lewis G, Gilbody S, Kessler D, Araya R, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. *Psychol Med*. 2022;52(10):1875-82.
373. Lemay KR, Tulloch HE, Pipe AL, Reed JL. Establishing the Minimal Clinically Important Difference for the Hospital Anxiety and Depression Scale in Patients With Cardiovascular Disease. *Journal of Cardiopulmonary Rehabilitation and Prevention*. 2019;39(6):E6-E11.
374. de Filippis R, Mercurio M, Segura-Garcia C, De Fazio P, Gasparini G, Galasso O. Defining the minimum clinically important difference (MCID) in the hospital anxiety and depression scale (HADS) in patients undergoing total hip and knee arthroplasty. *Orthopaedics & Traumatology: Surgery & Research*. 2023;103689.
375. Longo UG, Papalia R, De Salvatore S, Marinozzi A, Piergentili I, Lalli A, et al. Establishing the Minimum Clinically Significant Difference (MCID) and the Patient Acceptable Symptom Score (PASS) for the Hospital Anxiety and Depression Scale (HADS) in Patients with Rotator Cuff Disease and Shoulder Prosthesis. *Journal of Clinical Medicine*. 2023;12(4):1540.
376. Puhan MA, Frey M, Büchi S, Schünemann HJ. The minimal important difference of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease. *Health and Quality of Life Outcomes*. 2008;6(1):46.
377. Zhu M, Hong RH, Yang T, Yang X, Wang X, Liu J, et al. The Efficacy of Measurement-Based Care for Depressive Disorders: Systematic Review and Meta-Analysis of Randomized Controlled Trials. *J Clin Psychiatry*. 2021;82(5).

378. Zhou J, Wang X, Yang J, Zhu X, Xiao L, Feng L, et al. Optimization of measurement-based care (OMBC) for depression based on all-round and continuous assessment: rationale and protocol for a multicenter randomized control clinical trial. *Trials*. 2022;23(1):367.
379. Korsen N, Gerrish S. Use of PHQ-9 for monitoring patients with depression in integrated primary care practices. *Ann Fam Med*. 2022;20(20 Suppl 1).
380. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *Bmj*. 2019;365:l1476.
381. Narayanan S. The preanalytic phase. An important component of laboratory medicine. *Am J Clin Pathol*. 2000;113(3):429-52.
382. Lukmanji A, Williams JVA, Bulloch AGM, Bhattarai A, Patten SB. Seasonal variation in symptoms of depression: A Canadian population based study. *J Affect Disord*. 2019;255:142-9.
383. Gough A, Marshall T, Ferris E, Sitch A. Within-individual variation of measured depression symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders Reports*. 2023;14:100675.
384. Smarr KL, Keefer AL. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). *Arthritis Care Res (Hoboken)*. 2011;63(S11):S454-S66.
385. Kauffman KM, Dolata J, Figueroa M, Gunzler D, Huml A, Pencak J, et al. Directly Observed Weekly Fluoxetine for Major Depressive Disorder Among Hemodialysis Patients: A Single-Arm Feasibility Trial. *Kidney Med*. 2022;4(3):100413.
386. Bargi G, Bosnak Guclu M, Turkoz Sucak AG. Differences in pulmonary and extra-pulmonary characteristics in severely versus non-severely fatigued recipients of allogeneic hematopoietic stem cell transplantation: a cross-sectional, comparative study. *Hematol*. 2019;24(1):112-22.
387. Yang Q, Li M, Cao X, Lu Y, Tian C, Sun M, et al. An umbrella review of meta-analyses on diagnostic accuracy of C-reactive protein. *International Journal of Surgery*. 2022;104:106788.
388. Radicioni G, Ceppe A, Ford AA, Alexis NE, Barr RG, Bleecker ER, et al. Airway mucin MUC5AC and MUC5B concentrations and the initiation and progression of chronic obstructive pulmonary disease: an analysis of the SPIROMICS cohort. *Lancet Respir Med*. 2021;9(11):1241-54.
389. Maselli DJ, Bhatt SP, Anzueto A, Bowler RP, DeMeo DL, Diaz AA, et al. Clinical Epidemiology of COPD: Insights From 10 Years of the COPD Gene Study. *Chest*. 2019;156(2):228-38.
390. Ridker PM, Koenig W, Kastelein JJ, Mach F, Lüscher TF. Has the time finally come to measure hsCRP universally in primary and secondary cardiovascular prevention? *European Heart Journal*. 2018;39(46):4109-11.
391. Flynn R, Hedenmalm K, Murray-Thomas T, Pacurariu A, Arlett P, Shepherd H, et al. Ability of Primary Care Health Databases to Assess Medicinal Products Discussed by the European Union Pharmacovigilance Risk Assessment Committee. *Clinical Pharmacology & Therapeutics*. 2020;107(4):957-65.
392. Pélabon C, Hilde CH, Einum S, Gamelon M. On the use of the coefficient of variation to quantify and compare trait variation. *Evolution Letters*. 2020;4(3):180-8.
393. Kader GD, Perry M. Variability for Categorical Variables. *Journal of Statistics Education*. 2007;15(2):null-null.
394. Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. *SAGE Open Medicine*. 2020;8:2050312120934839.
395. Dolan E, O'Brien E. Blood Pressure Variability. *Hypertension*. 2010;56(2):179-81.

396. Chen Y, Ma Y, Qin J, Wei X, Yang Y, Yuan Y, et al. Blood pressure variability predicts poor outcomes in acute stroke patients without thrombolysis: a systematic review and meta-analysis. *J Neurol*. 2024;271(3):1160-9.
397. Chiu TJ, Yeh JT, Huang CJ, Chiang CE, Sung SH, Chen CH, et al. Blood pressure variability and cognitive dysfunction: A systematic review and meta-analysis of longitudinal cohort studies. *J Clin Hypertens (Greenwich)*. 2021;23(8):1463-82.
398. Musini VM, Wright JM. Factors affecting blood pressure variability: lessons learned from two systematic reviews of randomized controlled trials. *PLoS One*. 2009;4(5):e5673.
399. Bell K, Doust J, McGeechan K, Horvath AR, Barratt A, Hayen A, et al. The potential for overdiagnosis and underdiagnosis because of blood pressure variability: a comparison of the 2017 ACC/AHA, 2018 ESC/ESH and 2019 NICE hypertension guidelines. *Journal of Hypertension*. 9000;Publish Ahead of Print.
400. Zhu Y, Lu JM, Yu ZB, Li D, Wu MY, Shen P, et al. Intra-individual variability of total cholesterol is associated with cardiovascular disease mortality: A cohort study. *Nutr Metab Cardiovasc Dis*. 2019;29(11):1205-13.
401. Yan Y, Huang Y, Zhou D, Tang S, Feng YQ. Visit-to-Visit Variability in Total Cholesterol Correlates with the Progression of Renal Function Decline in a Chinese Community-Based Hypertensive Population. *Kidney & Blood Pressure Research*. 2019;44(4):727-42.
402. Snedecor GW. BIOLOGICAL VARIATION vs. ERRORS IN MEASUREMENT. *Science*. 1934;80(2072):246-7.
403. Sandberg S, Carobene A, Bartlett B, Coskun A, Fernandez-Calle P, Jonker N, et al. Biological variation: recent development and future challenges. *Clin Chem Lab Med*. 2023;61(5):741-50.
404. [NICE] NifHaCE. Suspected acute respiratory infection in over 16s: assessment at first presentation and initial management 2023 [Available from: <https://www.nice.org.uk/guidance/ng237/chapter/Recommendations>.
405. White S, Gong H, Zhu L, Doust J, Loh TP, Lord S, et al. Simulations found within-subject measurement variation in glycaemic measures may cause overdiagnosis of prediabetes and diabetes. *J Clin Epidemiol*. 2022;145:20-8.
406. Gottfried ED, Christopher SC. Mental Disorders Among Criminal Offenders: A Review of the Literature. *J Correct Health Care*. 2017;23(3):336-46.
407. Bell KJ, Glasziou PP, Hayen A, Irwig L. Criteria for monitoring tests were described: validity, responsiveness, detectability of long-term change, and practicality. *J Clin Epidemiol*. 2014;67(2):152-9.
408. Weykamp C, Siebelder C, Lenters E, Slingerland R, English E. The risk of clinical misinterpretation of HbA1c: Modelling the impact of biological variation and analytical performance on HbA1c used for diagnosis and monitoring of diabetes. *Clin Chim Acta*. 2023;548:117495.
409. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*. 1989;10(4):407-15.
410. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*. 2011;11(2):171-84.
411. Draak THP, de Greef BTA, Faber CG, Merckies ISJ. The minimum clinically important difference: which direction to take. *Eur J Neurol*. 2019;26(6):850-5.
412. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. 2003;41(5):582-92.
413. Hayen A, Bell K, Glasziou P, Irwig L. A counterargument to encounter frequency and target achievement: measurement variability. *Arch Intern Med*. 2012;172(4):374; author reply - 5.
414. Hall KH. Reviewing intuitive decision-making and uncertainty: the implications for medical education. *Medical Education*. 2002;36(3):216-24.

415. Cohen S, Cohen N, Gabbay E. Paternalism and certitude. *Bioethics*. 2020;34(5):478-82.
416. Benzel E. Paternalism versus Shared Decision Making. *World Neurosurg*. 2023;174:x.
417. Whiting PF, Davenport C, Jameson C, Burke M, Sterne JA, Hyde C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open*. 2015;5(7):e008155.
418. Stevens RJ, Oke J, Perera R. Statistical models for the control phase of clinical monitoring. *Statistical Methods in Medical Research*. 2010;19(4):394-414.
419. Ko SQ, Quah P, Lahiri M. The cost of repetitive laboratory testing for chronic disease. *Intern Med J*. 2019;49(9):1168-70.
420. Vrijnsen BEL, Naaktgeboren CA, Vos LM, van Solinge WW, Kaasjager HAH, Ten Berg MJ. Inappropriate laboratory testing in internal medicine inpatients: Prevalence, causes and interventions. *Ann Med Surg (Lond)*. 2020;51:48-53.
421. van Walraven C, Raymond M. Population-based study of repeat laboratory testing. *Clin Chem*. 2003;49(12):1997-2005.
422. Monach PA. Repeating tests: Different roles in research studies and clinical medicine. *Biomarkers in Medicine*. 2012;6(5):691-703.
423. Thompson HK, Woodbury MA. Clinical data representation in multidimensional space. *Computers and Biomedical Research*. 1970;3(1):58-73.
424. Bauer DT, Guerlain S, Brown PJ. The design and evaluation of a graphical display for laboratory data. *J Am Med Inform Assoc*. 2010;17(4):416-24.
425. Coskun A, Lippi G. Personalized laboratory medicine in the digital health era: recent developments and future challenges. *Clin Chem Lab Med*. 2024;62(3):402-9.
426. Coskun A, Lippi G. The impact of physiological variations on personalized reference intervals and decision limits: an in-depth analysis. *Clin Chem Lab Med*. 2024.
427. German CA, Sinsheimer JS, Zhou J, Zhou H. WiSER: Robust and scalable estimation and inference of within-subject variances from intensive longitudinal data. *Biometrics*. 2022;78(4):1313-27.
428. Yeh ST, Ooi SW, Chang YH, Li CY, Chen HF. Age and sex-specific associations of visit-to-visit variability of glycated hemoglobin A1c with all-cause mortality in patients with diabetes. *Medicine (Baltimore)*. 2022;101(33):e29942.
429. Agardh CD, Agardh E, Torffvit O. The association between retinopathy, nephropathy, cardiovascular disease and long-term metabolic control in type 1 diabetes mellitus: a 5 year follow-up study of 442 adult patients in routine care. *Diabetes Research & Clinical Practice*. 1997;35(2-3):113-21.
430. Barton K, Donovan L, Giroux I, Miller D, Mottola M, McManus R. Glycated hemoglobin measurements at three, 12 and 24 months postpartum after gestational diabetes. *Clinical and investigative medicine*. 2019;Medecine clinique et experimentale. 42(4):E37-E41.
431. Borsey DQ, Fraser DM, Gray RS. Glycosylated hemoglobin and its temporal relationship to plasma glucose in non-insulin dependent (type 2) diabetes mellitus. *Metabolism: Clinical and Experimental*. 1982;31(4):362-5.
432. Bouchi R, Babazono T, Mugishima M, Yoshida N, Nyumura I, Toya K, et al. Fluctuations in HbA1c are associated with a higher incidence of cardiovascular disease in Japanese patients with type 2 diabetes. *Journal of Diabetes Investigation*. 2012;3(2):148-55.
433. Braga F, Dolci A, Montagnana M, Pagani F, Paleari R, Guidi GC, et al. Revaluation of biological variation of glycated hemoglobin (HbA1c) using an accurately designed protocol and an assay traceable to the IFCC reference system. *Clinica Chimica Acta*. 2011;412(15-16):1412-6.
434. Cardoso CRL, Leite NC, Moram CBM, Salles GF. Long-term visit-to-visit glycemic variability as predictor of micro- and macrovascular complications in patients with type 2 diabetes: The Rio de Janeiro Type 2 Diabetes Cohort Study. *Cardiovascular Diabetology*. 2018;17(1).

435. Ceriello A, De Cosmo S, Rossi MC, Lucisano G, Genovese S, Pontremoli R, et al. Variability in HbA1c, blood pressure, lipid parameters and serum uric acid, and risk of development of chronic kidney disease in type 2 diabetes. *Diabetes Obes Metab*. 2017;19(11):1570-8.
436. Chiu HT, Li TC, Li CI, Liu CS, Lin WY, Lin CC. Visit-to-visit glycemic variability is a strong predictor of chronic obstructive pulmonary disease in patients with type 2 diabetes mellitus: Competing risk analysis using a national cohort from the Taiwan diabetes study. *PLoS ONE*. 2017;12(5).
437. Coskun A, Carobene A, Kilercik M, Serteser M, Sandberg S, Aarsand AK, et al. Within-subject and between-subject biological variation estimates of 21 hematological parameters in 30 healthy subjects. *Clinical Chemistry and Laboratory Medicine*. 2018;56(8):1309-18.
438. Critchley JA, Carey IM, Harris T, DeWilde S, Cook DG. Variability in glycated hemoglobin and risk of poor outcomes among people with type 2 diabetes in a large primary care cohort study. *Diabetes Care*. 2019;42(12):2237-46.
439. Dergez E, Lanyi E. Calculation of a reference change value with application of locally estimated biological variation of Hemoglobin A1c. *Clinical Chemistry and Laboratory Medicine*. 2014;52(9):eA97.
440. Desmeules P, Cousineau J, Allard P. Biological variation of glycated haemoglobin in a paediatric population and its application to calculation of significant change between results. *Ann Clin Biochem*. 2010;47(Pt 1):35-8.
441. Dorajoo SR, Ng JSL, Goh JHF, Lim SC, Yap CW, Chan A, et al. HbA1c variability in type 2 diabetes is associated with the occurrence of new-onset albuminuria within three years. *Diabetes Research and Clinical Practice*. 2017;128:32-9.
442. Ercan M, Akbulut D, Oguz E, Bal C, Cicek EA, Yilmaz FM. Estimation of biological variation components of glycated hemoglobin HbA1c on Tosoh G7 HPLC. *Turkish Journal of Biochemistry*. 2015;40 (Supplement 1):157.
443. Ercan M, Avci E, Akbulut ED, Oguz EF, Topcuoglu C, Yucel C, et al. Biological variation components of glycated hemoglobin HbA1C on tosoh G7 HPLC: Biological variation components of HbA1C. *Journal of Clinical and Analytical Medicine*. 2018;9(6):469-73.
444. Foo V, Quah J, Cheung G, Tan NC, Ma Zar KL, Chan CM, et al. HbA1c, systolic blood pressure variability and diabetic retinopathy in Asian type 2 diabetics. *J Diabetes*. 2017;9(2):200-7.
445. Godsland IF. Intra-individual variation: significant changes in parameters of lipid and carbohydrate metabolism in the individual and intra-individual variation in different test populations. *Annals of Clinical Biochemistry*. 1985;22(Pt 6):618-24.
446. Gu J, Fan YQ, Zhang JF, Wang CQ. Association of hemoglobin A1c variability and the incidence of heart failure with preserved ejection fraction in patients with type 2 diabetes mellitus and arterial hypertension. *Hellenic Journal of Cardiology*. 2018;59(2):91-7.
447. Gu J, Fan YQ, Zhang JF, Wang CQ. Impact of long-term glycemic variability on development of atrial fibrillation in type 2 diabetic patients. *Anatolian Journal of Cardiology*. 2017;18(6):410-6.
448. Hashimoto Y, Kaji A, Sakai R, Osaka T, Ushigome E, Hamaguchi M, et al. Skipping breakfast is associated with glycemic variability in patients with type 2 diabetes. *Nutrition*. 2020;71 (no pagination).
449. He H, Peng S, Liang L, Zhou J, An Z, Huang H. Biological variation and reference change value of HbA1c and glycated albumin in healthy individuals during short-term follow-up. *Clinical Chemistry and Laboratory Medicine*. 2015;53:S603.
450. Herath M, Weeraratna TP. Lack of association between different metrics of glycemic control with progression of nephropathy in type 2 diabetes: A preliminary observation from a Sri Lankan cohort. *Diabetes*. 2016;65 (Supplement 1):A550.

451. Hermann JM, Hammes H-P, Rami-Merhar B, Rosenbauer J, Schütt M, Siegel E, et al. HbA1c Variability as an Independent Risk Factor for Diabetic Retinopathy in Type 1 Diabetes: A German/Austrian Multicenter Analysis on 35,891 Patients. *PLOS ONE*. 2014;9(3):e91137.
452. Hietala K, Wadén J, Forsblom C, Harjutsalo V, Kytö J, Summanen P, et al. HbA1c variability is associated with an increased risk of retinopathy requiring laser treatment in type 1 diabetes. *Diabetologia*. 2013;56(4):737-45.
453. Hirakawa Y, Arima H, Zoungas S, Ninomiya T, Cooper M, Hamet P, et al. Impact of visit-to-visit glycemic variability on the risks of macrovascular and microvascular events and all-cause mortality in type 2 diabetes: The ADVANCE trial. *Diabetes Care*. 2014;37(8):2359-65.
454. Howey JE, Bennet WM, Browning MC, Jung RT, Fraser CG. Clinical utility of assays of glycosylated haemoglobin and serum fructosamine compared: use of data on biological variation. *Diabetic Medicine*. 1989;6(9):793-6.
455. Hsu CC, Chang HY, Huang MC, Hwang SJ, Yang YC, Lee YS, et al. HbA1c variability is associated with microalbuminuria development in type 2 diabetes: a 7-year prospective cohort study. *Diabetologia*. 2012;55(12):3163-72.
456. Jang G, Jeong K, Kim J, Song TW, Min TK, Kim WK, et al. Post traumatic stress disorder (PTSD) in parents of pediatric anaphylaxis patients; using BDI-II, BAI, And IES-R. *Allergy: European Journal of Allergy and Clinical Immunology*. 2019;74 (Supplement 106):592-3.
457. Jun JE, Jin SM, Baek J, Oh S, Hur KY, Lee MS, et al. The association between glycemic variability and diabetic cardiovascular autonomic neuropathy in patients with type 2 diabetes. *Cardiovascular Diabetology*. 2015;14(1).
458. Jury DR, Baker JR, Dunn PJ. Clinical importance of the reversible fraction of haemoglobin A1c in Type 2 (non-insulin-dependent) diabetes. *Diabetologia*. 1983;25(4):313-5.
459. Kilpatrick ES, Maylor PW, Keevil BG. Biological variation of glycated hemoglobin. Implications for diabetes screening and monitoring. *Diabetes Care*. 1998;21(2):261-4.
460. Kilpatrick ES, Rigby AS, Atkin SL. A1C variability and the risk of microvascular complications in type 1 diabetes: data from the Diabetes Control and Complications Trial. *Diabetes Care*. 2008;31(11):2198-202.
461. Kitaoka K, Takenouchi A, Tsuboi A, Fukuo K, Kazumi T. Association of Postbreakfast Triglyceride and Visit-to-Visit Annual Variation of Fasting Plasma Glucose with Progression of Diabetic Nephropathy in Patients with Type 2 Diabetes. *Journal of Diabetes Research*. 2016;2016:4351376.
462. Lee CL, Chen CH, Wu MJ, Tsai SF. The variability of glycated hemoglobin is associated with renal function decline in patients with type 2 diabetes. *Therapeutic Advances in Chronic Disease*. 2020;11(no pagination).
463. Lenters-Westra E, Roraas T, Schindhelm RK, Slingerland RJ, Sandberg S. Biological variation of hemoglobin A1c: consequences for diagnosing diabetes mellitus. *Clinical Chemistry*. 2014;60(12):1570-2.
464. Li TC, Yang CP, Tseng ST, Li CI, Liu CS, Lin WY, et al. Visit-to-visit variations in fasting plasma glucose and HbA1c associated with an increased risk of Alzheimer disease: Taiwan diabetes study. *Diabetes Care*. 2017;40(9):1210-7.
465. Lin CC, Chen CC, Chen FN, Li CI, Liu CS, Lin WY, et al. Risks of diabetic nephropathy with variation in hemoglobin A<inf>1c</inf> and fasting plasma glucose. *American Journal of Medicine*. 2013;126(11):1017.e1-.e10.
466. Low S, Lim SC, Yeoh LY, Liu YL, Liu JJ, Fun S, et al. Effect of long-term glycemic variability on estimated glomerular filtration rate decline among patients with type 2 diabetes mellitus: Insights from the Diabetic Nephropathy Cohort in Singapore. *Journal of Diabetes*. 2017;9(10):908-19.
467. Luk AOY, Ma RCW, Lau ESH, Yang X, Lau WWY, Yu LWL, et al. Risk association of HbA1c variability with chronic kidney disease and cardiovascular disease in type 2 diabetes:

- prospective analysis of the Hong Kong Diabetes Registry. *Diabetes Metab Res Rev*. 2013;29(5):384-90.
468. Ma CJ, Sun LC, Chen FM, Lu CY, Shih YL, Tsai HL, et al. A double-blind randomized study comparing the efficacy and safety of a composite vs a conventional intravenous fat emulsion in postsurgical gastrointestinal tumor patients. *Nutrition in Clinical Practice*. 2012;27(3):410-5.
469. Mao Y, Zhong W. HbA1c Variability as an Independent Risk Factor for Microvascular Complications in Type 1 Diabetes. *Journal of Diabetes Science and Technology*. 2022.
470. Matsutani D, Sakamoto M, Iuchi H, Minato S, Suzuki H, Kayama Y, et al. Glycemic variability in continuous glucose monitoring is inversely associated with baroreflex sensitivity in type 2 diabetes: A preliminary report. *Cardiovascular Diabetology*. 2018;17(1).
471. Meigs JB, Nathan DM, Cupples LA, Wilson PWF, Singer DE. Tracking of glycated hemoglobin in the original cohort of the Framingham Heart Study. *Journal of Clinical Epidemiology*. 1996;49(4):411-7.
472. Minami T, Ito Y, Yamada M, Furuta R, Minagawa F, Kamata K, et al. The effect of long-term past glycemic control on executive function among patients with type 2 diabetes mellitus. *Diabetology International*. 2020;11(2):114-20.
473. Miyamoto H, Suzuki H, Yokoyama Y, Akizuki S, Hirai N, Ohnishi A. Relationship between glycated albumin (GA) and glycated hemoglobin (A1c) in 255 patients with liver diseases using cross-sectional laboratory data. [Japanese]. *Rinsho byori*. 2008;The Japanese journal of clinical pathology. 56(9):761-6.
474. Moon D, Chung WS, Yoo KD. Variability in haemoglobin A1c predicts clinical outcomes in patients with asymptomatic type 2 diabetes. *EuroIntervention*. 2016;France.:315.
475. Nagl K, Gericke G, Duffles E, Fritsch M, Hortenhuber T, Rami-Merhar B, et al. Endothelial progenitor cells show relation to asymmetric dimethylarginine (ADMA) levels and HbA1c variability in young patients with type 1 without complications. *Diabetologia*. 2014;57(1):S526.
476. Nazim J, Fendler W, Starzyk J. Metabolic control and its variability are major risk factors for microalbuminuria in children with type 1 diabetes. *Endokrynologia Polska*. 2014;65(2):83-9.
477. Nichols GA, Gandra SR, Chiou CF, Anthony MS, Alexander-Bridges M, Brown JB. Successes and challenges of insulin therapy for type 2 diabetes in a managed-care setting. *Current Medical Research & Opinion*. 2010;26(1):9-15.
478. Molinaro RJ, Noguez JH. Hb A1c Testing at Proper Frequencies Does not Depict Appropriate Changes in Treatment. *American Journal of Clinical Pathology*. 2013;140(suppl_1):A121-A.
479. Parkes J, Ray R, Kerestan S, Davis H, Ginsberg B. Prospective evaluation of accuracy, precision, and reproducibility of an at-home hemoglobin A1c sampling kit. *Diabetes technology & therapeutics*. 1999;1(4):411-9.
480. Parrinello CM, Lutsey P, Couper D, Eckfeldt JH, Coresh J, Selvin E. Short-term total variability in nontraditional biomarkers of hyperglycemia in older adults. *Diabetes*. 2015;64:A433-A4.
481. Pecoraro RE, Koepsell TD, Chen MS. Comparative clinical reliability of fasting plasma glucose and glycosylated hemoglobin in non-insulin-dependent diabetes mellitus. *Diabetes Care*. 1986;9(4):370-5.
482. Penno G, Solini A, Zoppini G, Orsi E, Fondelli C, Zerbini G, et al. Hemoglobin A1c variability as an independent correlate of cardiovascular disease in patients with type 2 diabetes: a cross-sectional analysis of the Renal Insufficiency and Cardiovascular Events (RIACE) Italian Multicenter Study. *Cardiovascular Diabetology*. 2013;12(1):98.
483. Phillipou G, Phillips PJ. Intraindividual variation of glycohemoglobin: implications for interpretation and analytical goals. *Clinical Chemistry*. 1993;39(11 Pt 1):2305-8.

484. Pimentel AL, Camargo JL. Variability of glycated hemoglobin levels in the first year post renal transplantation in patients without diabetes. *Clinical Biochemistry*. 2017;50(18):997-1001.
485. Prentice JC, Pizer SD, Conlin PR. Identifying the independent effect of HbA_{1c} variability on adverse health outcomes in patients with Type 2 diabetes. *Diabetic Medicine*. 2016;33(12):1640-8.
486. Ramachandran A, Winter M, Mann DM. Association of visit-to-visit variability of hemoglobin A1c and medication adherence. *Journal of Managed Care & Specialty Pharmacy*. 2015;21(3):229-37.
487. Rodriguez-Segade S, Rodriguez J, Garcia Lopez JM, Casanueva FF, Camina F. Estimation of the glycation gap in diabetic patients with stable glycemic control. *Diabetes Care*. 2012;35(12):2447-50.
488. Rohlfing C, Wiedmeyer HM, Little R, Grotz VL, Tennill A, England J, et al. Biological variation of glycohemoglobin. *Clinical Chemistry*. 2002;48(7):1116-8.
489. Romero-Aroca P, Navarro-Gil R, Feliu A, Valls A, Moreno A, Baget-Bernaldiz M. The Effect of HbA_{1c} Variability as a Risk Measure for Microangiopathy in Type 1 Diabetes Mellitus. *Diagnostics*. 2021;11(7):24.
490. Rosa LCGFD, Zajdenverg L, Souto DL, Dantas JR, Pinto MVR, Salles GFDCMD, et al. HbA_{1c} variability and long-term glycemic control are linked to diabetic retinopathy and glomerular filtration rate in patients with type 1 diabetes and multiethnic background. *Journal of Diabetes and its Complications*. 2019;33(9):610-5.
491. Rutter CE, Millard LAC, Borges MC, Lawlor DA. Exploring regression dilution bias using repeat measurements of 2858 variables in up to 49 000 UK Biobank participants. *medRxiv*. 2022;15.
492. Sakai R, Hashimoto Y, Hamaguchi M, Ushigome E, Okamura T, Majima S, et al. Living alone is associated with visit-to-visit hba1c variability in men but not in women in people with type 2 diabetes: Kamogawa-dm cohort study. *Endocrine Journal*. 2020;67(4):419-26.
493. Sambataro M, Tura A, Sambado L, Trevisiol E, Seganfredo E, Durante E, et al. Possible clinical prognostic meaning of glucose variability indexes in type 2 diabetes (T2D) with lower-limb neuropathic and neuroischemic lesions-Italian leukemia association treviso section project. *Diabetes*. 2017;66 (Supplement 1):A590-A1.
494. Savu O, Elian V, Steriade O, Teodoru I, Mihut S, Tacu C, et al. The impact of basal insulin analogues on glucose variability in patients with type 2 diabetes undergoing renal replacement therapy for end-stage renal disease. *International Urology & Nephrology*. 2016;48(2):265-70.
495. Schreur V, van Asten F, Ng H, Weeda J, Groenewoud JMM, Tack CJ, et al. Risk factors for development and progression of diabetic retinopathy in Dutch patients with type 1 diabetes mellitus. *Acta Ophthalmologica*. 2018;96(5):459-64.
496. Scott ES, Januszewski AS, O'Connell R, Colagiuri S, Fulcher GR, Keech A, et al. HbA_{1c} coefficient of variation is an independent risk factor for chronic complications in type 2 diabetes in the FIELD study. *Diabetologia*. 2018;61 (Supplement 1):S519.
497. Scott ES, McGrath RT, Januszewski AS, Calandro D, Hardikar AA, O'Neal DN, et al. HbA_{1c} variability in adults with type 1 diabetes on continuous subcutaneous insulin infusion (CSII) therapy compared to multiple daily injection (MDI) treatment. *BMJ Open*. 2019;9(12).
498. Selvin E, Crainiceanu CM, Brancati FL, Coresh J. Short-term variability in measures of glycemia and implications for the classification of diabetes. *Archives of Internal Medicine*. 2007;167(14):1545-51.
499. Sheedy C, Kyaw Tun T, McDermott JH, Sreenan S. Visit-to-visit HbA_{1c} variability as a risk factor for retinopathy and microalbuminuria in people with diabetes. *Irish Journal of Medical Science*. 2022;191(Supplement 4):S120.
500. Su JB, Zhao LH, Zhang XL, Cai HL, Huang HY, Xu F, et al. HbA_{1c} variability and diabetic peripheral neuropathy in type 2 diabetic patients. *Cardiovascular Diabetology*. 2018;17(1).

501. Sun B, He F, Gao Y, Zhou J, Sun L, Liu R, et al. Prognostic impact of visit-to-visit glycemic variability on the risks of major adverse cardiovascular outcomes and hypoglycemia in patients with different glycemic control and type 2 diabetes. *Endocrine*. 2019.
502. Takahashi O, Ohde S. The optimal interval of diabetes mellitus screening in healthy adults: A random effects model. *Value Health*. 2014;17 (3):A242.
503. Takao T, Matsuyama Y, Yanagisawa H, Kikuchi M, Kawazu S. Association between HbA1c variability and mortality in patients with type 2 diabetes. *Journal of Diabetes and its Complications*. 2014;28(4):494-9.
504. Takenouchi A, Tsuboi A, Kurata M, Fukuo K, Kazumi T. Carotid Intima-Media Thickness and Visit-to-Visit HbA1c Variability Predict Progression of Chronic Kidney Disease in Type 2 Diabetic Patients with Preserved Kidney Function. *Journal of Diabetes Research*. 2016;2016 (no pagination).
505. Teliti M, Cogni G, Sacchi L, Dagliati A, Marini S, Tibollo V, et al. Risk factors for the development of micro-vascular complications of type 2 diabetes in a single-centre cohort of patients. *Diabetes & Vascular Disease Research*. 2018;15(5):424-32.
506. Thyagarajan B, Howard AG, Durazo-Arvizu R, Eckfeldt JH, Gellman MD, Kim RS, et al. Analytical and biological variability in biomarker measurement in the Hispanic Community Health Study/Study of Latinos. *Clinica Chimica Acta*. 2016;463:129-37.
507. Tjaden AH, Edelstein S, Arslanian SA, Barengolts E, Caprio S, Cree-Green M, et al. Reproducibility of glycemic measures among youth and adults with impaired glucose tolerance (IGT) or recently diagnosed type 2 diabetes (T2D) in the restoring insulin secretion (RISE) study. *Diabetes Conference: 79th Scientific Sessions of the American Diabetes Association, ADA*. 2019;68(Supplement 1).
508. Trape J, Aliart ML, Brunet M, Dern E, Abadal E, Queralto JM. Reference change value for HbA1c in patients with type 2 diabetes mellitus. *Clinical Chemistry & Laboratory Medicine*. 2000;38(12):1283-7.
509. Tseng JY, Chen HH, Huang KC, Hsu SP, Chen CC. Effect of mean HbA1c on the association of HbA1c variability and all-cause mortality in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism*. 2020;22(4):680-7.
510. Tsuboi A, Takenouchi A, Kurata M, Fukuo K, Kazumi T. Postmeal triglyceridemia and variability of HbA1c and postmeal glycemia were predictors of annual decline in estimated glomerular filtration rate in type 2 diabetic patients with different stages of nephropathy. *Journal of Diabetes and Metabolic Disorders*. 2017;16(1).
511. Tsuchiya T, Saisho Y, Murakami R, Watanabe Y, Inaishi J, Itoh H. Relationship between daily and visit-to-visit glycemic variability in patients with type 2 diabetes. *Endocrine journal*. 2020;08.
512. Tynjala A, Forsblom C, Harjutsalo V, Groop PH, Gordin D. HbA1c variability is associated with arterial stiffness in type 1 diabetes. *Diabetologia*. 2020;63(SUPPL 1):S441-S2.
513. Ucar F, Erden G, Ginis Z, Ozturk G, Sezer S, Gurler M, et al. Estimation of biological variation and reference change value of glycated hemoglobin (HbA1c) when two analytical methods are used. *Clinical Chemistry*. 2013;59(10):A87.
514. Cassiani-Miranda CA, Cuadros-Cruz AK, Torres-Pinzon H, Scoppetta O, Pinzon-Tarrazona JH, Lopez-Fuentes WY, et al. Validity of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adult primary care users in Bucaramanga, Colombia. *Revista Colombiana de Psiquiatria*. 2020.
515. Waden J, Forsblom C, Thorn LM, Gordin D, Saraheimo M, Groop P-H. A1C variability predicts incident cardiovascular events, microalbuminuria, and overt diabetic nephropathy in patients with type 1 diabetes. *Diabetes*. 2009;58(11):2649.
516. Walker GS, Cunningham SG, Sainsbury CAR, Jones GC. HbA1c variability is associated with increased mortality and earlier hospital admission in people with Type 1 diabetes. *Diabetic Medicine*. 2017;34(11):1541-5.

517. Watts GF, Morris RW, Goodland FC, Kubal C, Shaw KM. Serum fructosamine and glycosylated haemoglobin in the monitoring of glycaemic control in insulin-dependent diabetic outpatients. *Practical Diabetes*. 1989;6(4):159-63.
518. Whitsel EA, Nguyen QC, C MS, Tabor JW, Cuthbertson CC, Wener MH, et al. Dried capillary whole blood spot-based hemoglobin A1C, fasting glucose, and diabetes prevalence in a nationally representative population of young U.S. adults: Add health, wave IV. *Circulation Conference: Epidemiology and Prevention/Physical Activity, Nutrition and Metabolism*. 2012;125(10 SUPPL. 1).
519. Wightman S, Sainsbury CAR, Jones GC. HbA1c variability and systolic blood pressure variability confer significant and additive risk of mortality in people with Type 1 diabetes. *Diabetic Medicine*. 2018;35 (Supplement 1):77.
520. Yang CY, Su PF, Hung JY, Ou HT, Kuo S. Comparative predictive ability of visit-to-visit HbA1c variability measures for microvascular disease risk in type 2 diabetes. *Cardiovascular Diabetology*. 2020;19(1).
521. Yang Y, Lee EY, Cho JH, Park YM, Ko SH, Yoon KH, et al. Cardiovascular autonomic neuropathy predicts higher hba1c variability in subjects with type 2 diabetes mellitus. *Diabetes and Metabolism Journal*. 2018;42(6):496-512.
522. Yang YF, Li TC, Li CI, Liu CS, Lin WY, Yang SY, et al. Visit-to-Visit Glucose Variability Predicts the Development of End-Stage Renal Disease in Type 2 Diabetes: 10-Year Follow-Up of Taiwan Diabetes Study. *Medicine*. 2015;94(44):e1804.
523. Yang HK, Kang B, Lee SH, Yoon KH, Hwang BH, Chang K, et al. Association between hemoglobin A1c variability and subclinical coronary atherosclerosis in subjects with type 2 diabetes. *Journal of Diabetes & its Complications*. 2015;29(6):776-82.
524. Zhong VW, Juhaeri J, Cole SR, Shay CM, Gordon-Larsen P, Kontopantelis E, et al. HbA1C variability and hypoglycemia hospitalization in adults with type 1 and type 2 diabetes: A nested case-control study. *Journal of diabetes and its complications*. 2018;32(2):203-9.
525. Alexander KS, Kazmierczak SC, Snyder CK, Oberdorf JA, Farrell DH. Prognostic utility of biochemical markers of cardiovascular risk: impact of biological variability. *Clinical Chemistry & Laboratory Medicine*. 2013;51(9):1875-82.
526. Bailey D, Bevilacqua V, Colantonio DA, Pasic MD, Perumal N, Chan MK, et al. Pediatric within-day biological variation and quality specifications for 38 biochemical markers in the CALIPER cohort. *Clin Chem*. 2014;60(3):518-29.
527. Bay-Jensen AC, Bihlet A, Byrjalsen I, Andersen J, He Y, Siebuhr A, et al. Elevated levels of CRPM, an inflammatory biomarker correlating with disease activity in RA, are prognostic of radiographic knee OA. *Osteoarthritis and Cartilage*. 2017;25 (Supplement 1):S32.
528. Becker M, Sperlich B, Zinner C, Achtzehn S. Intra-Individual and Seasonal Variation of Selected Biomarkers for Internal Load Monitoring in U-19 Soccer Players. *Front Physiol*. 2020;11:838.
529. Block G, Dietrich M, Norkus E, Jensen C, Benowitz N, L., Morrow J, D. , et al. Intraindividual Variability of Plasma Antioxidants, Markers of Oxidative Stress, C-Reactive Protein, Cotinine, and Other Biomarkers. *Epidemiology*. 2006;17(4):404-12.
530. Blum CA, Muller B, Huber P, Kraenzlin M, Schindler C, De Geyter C, et al. Low-grade inflammation and estimates of insulin resistance during the menstrual cycle in lean and overweight women. *Journal of Clinical Endocrinology and Metabolism*. 2005;90(6):3230-5.
531. Boenisch O, Ehmke KD, Heddergott A, Naoum C, Frei U, Schindler R. C-reactive-protein and cytokine plasma levels in hemodialysis patients. *Journal of Nephrology*. 2002;15(5):547-51.
532. Bogaty P, Dagenais GR, Joseph L, Boyer L, Leblanc A, Belisle P, et al. Time variability of C-reactive protein: implications for clinical risk stratification. *PLoS ONE [Electronic Resource]*. 2013;8(4):e60759.

533. Bogaty P, Brophy JM, Boyer L, Simard S, Joseph L, Bertrand F, et al. Fluctuating Inflammatory Markers in Patients With Stable Ischemic Heart Disease. *Arch Intern Med*. 2005;165(2):221-6.
534. Browning LM, Jebb SA, Mishra GD, Cooke JH, O'Connell MA, Crook MA, et al. Elevated sialic acid, but not CRP, predicts features of the metabolic syndrome independently of BMI in women. *International Journal of Obesity & Related Metabolic Disorders: Journal of the International Association for the Study of Obesity*. 2004;28(8):1004-10.
535. Carobene A, Aarsand AK, Guerra E, Bartlett WA, Coskun A, Diaz-Garzon J, et al. European Biological Variation Study (EuBIVAS): Within- and Between-Subject Biological Variation Data for 15 Frequently Measured Proteins. *Clinical Chemistry*. 2019;65(8):1031-41.
536. Cho LW, Jayagopal V, Kilpatrick ES, Atkin SL. The biological variation of C-reactive protein in polycystic ovarian syndrome. *Clinical Chemistry*. 2005;51(10):1905-7.
537. Christmas B, Taylor L, Smith A, Pemberton P, Siegler JC, Midgley AW. Reproducibility of measurement techniques used for creatine kinase, interleukin-6 and high-sensitivity C-reactive protein determination over a 48 h period in males and females. *Measurement in Physical Education and Exercise Science*. 2018;22(3):191-9.
538. Clark GH, Fraser CG. Biological variation of acute phase proteins. *Annals of Clinical Biochemistry*. 1993;30(Pt 4):373-6.
539. Corte Z, Venta R. Biological variation of metabolic cardiovascular risk factors in haemodialysis patients and healthy individuals. *Annals of Translational Medicine*. 2020;8(6).
540. de Maat MPM, de Bart ACW, Hennis BC, Meijer P, Havelaar AC, Mulder PGH, et al. Interindividual and Intraindividual Variability in Plasma Fibrinogen, TPA Antigen, PAI Activity, and CRP in Healthy, Young Volunteers and Patients With Angina Pectoris. *Arterioscler Thromb Vasc Biol*. 1996;16(9):1156-62.
541. DeGoma EM, French B, Dunbar RL, Allison MA, Mohler ER, 3rd, Budoff MJ. Intraindividual variability of C-reactive protein: the Multi-Ethnic Study of Atherosclerosis. *Atherosclerosis*. 2012;224(1):274-9.
542. d'Eril GM, Anesi A, Maggiore M, Leoni V. Biological variation of serum amyloid A in healthy subjects. *Clinical Chemistry*. 2001;47(8):1498-9.
543. Engelberger RP, Limacher A, Kucher N, Baumann F, Silbernagel G, Benghozi R, et al. Biological variation of established and novel biomarkers for atherosclerosis: Results from a prospective, parallel-group cohort study. *Clinica Chimica Acta*. 2015;447:16-22.
544. Franzini C, Scapallato, L., Vanoni, G. Variabilita biologica ed analitica della anti-streptolisina O e di proteine della fase acuta del siero. *Biochimica Clinica*. 1993;17:315-9.
545. Gasco J, Inigo V, Mascaros V, Bernabeu R, Servera M. Biological variation of beta2-microglobulin in haemodiafiltration patients. *NDT Plus*. 2010;3:iii459.
546. Glynn RJ, MacFadyen JG, Ridker PM. Tracking of high-sensitivity C-reactive protein after an initially elevated concentration: The JUPITER study. *Clinical Chemistry*. 2009;55(2):305-12.
547. Guy JH, Edwards AM, Miller CM, Deakin GB, Pyne DB. Short-term reliability of inflammatory mediators and response to exercise in the heat. *Journal of sports sciences*. 2017;35(16):1622-8.
548. Hardikar S, Song X, Kratz M, Anderson GL, Blount PL, Reid BJ, et al. Intraindividual variability over time in plasma biomarkers of inflammation and effects of long-term storage. *Cancer Causes and Control*. 2014;25(8):969-76.
549. Hosogaya S, Naito K, Sakamoto M, Osada M, Yatomi Y, Ozaki Y. Biological inter- and intra-individual variations of serum immunochemical constituents and their allowable limits of analytical error. [Japanese]. *Rinsho byori*. 1999;The Japanese journal of clinical pathology. 47(9):875-80.
550. Itoh H, Mori I, Matsumoto Y, Maki S, Ogawa Y. Seasonal and inter-day variation in serum high-sensitivity C-reactive protein in Japanese male workers: A longitudinal study. *Industrial Health*. 2012;50(1):60-3.

551. Jackson SE, van Jaarsveld CH, Beeken RJ, Gunter MJ, Steptoe A, Wardle J. Four-year stability of anthropometric and cardio-metabolic parameters in a prospective cohort of older adults. *Biomarkers in Medicine*. 2015;9(2):109-22.
552. Jain P, Kadeangadi DM. Depression and anxiety among adult patients with type II diabetes mellitus- a descriptive study in Urban North Karnataka. *Journal of Clinical and Diagnostic Research*. 2020;14(6):LC11-LC4.
553. Jung D, Jang J, Kwong E, Song J, Quon B. Serum CRP and calprotectin to diagnose CF pulmonary exacerbations. *Pediatr Pulmonol*. 2019;54 (Supplement 2):353.
554. Koenig W, Sund M, Frohlich M, Lowel H, Hutchinson WL, Pepys MB. Refinement of the association of serum C-reactive protein concentration and coronary heart disease risk by correction for within-subject variation over time: The MONICA Augsburg studies, 1984 and 1987. *American Journal of Epidemiology*. 2003;158(4):357-64.
555. Kolsum U, Roy K, Starkey C, Borrill Z, Truman N, Vestbo J, et al. The repeatability of interleukin-6, tumor necrosis factor-alpha, and C-reactive protein in COPD patients over one year. *International Journal of Copd*. 2009;4:149-56.
556. Laclair R, O'Neal K, Ofner S, Sosa MJ, Labarrere CA, Moe SM. Precision of biomarkers to define chronic inflammation in CKD. *American Journal of Nephrology*. 2008;28(5):808-12.
557. Lee SA, Kallianpur A, Xiang YB, Wen W, Cai Q, Liu D, et al. Intra-individual variation of plasma adipokine levels and utility of single measurement of these biomarkers in population-based studies. *Cancer Epidemiology Biomarkers and Prevention*. 2007;16(11):2464-70.
558. Macy EM, Hayes TE, Tracy RP. Variability in the measurement of C-reactive protein in healthy subjects: implications for reference intervals and epidemiological applications. *Clinical Chemistry*. 1997;43(1):52-8.
559. Meyer T, Meister S. Routine blood parameters in elite soccer players. *International Journal of Sports Medicine*. 2011;32(11):875-81.
560. Nasermoaddeli A, Sekine M, Kagamimori S. Intra-individual variability of high-sensitivity C-reactive protein: age-related variations over time in Japanese subjects. *Circulation Journal*. 2006;70(5):559-63.
561. Navarro SL, Brasky TM, Schwarz Y, Song X, Wang CY, Kristal AR, et al. Reliability of serum biomarkers of inflammation from repeated measures in healthy individuals. *Cancer Epidemiology Biomarkers and Prevention*. 2012;21(7):1167-70.
562. Ockene IS, Matthews CE, Rifai N, Ridker PM, Reed G, Stanek E. Variability and Classification Accuracy of Serial High-Sensitivity C-Reactive Protein Measurements in Healthy Adults. *Clin Chem*. 2001;47(3):444-50.
563. Park S, Phillips JE, Saiphoklang N, Wilgus ML, Buhr RG, Cooper CB, et al. Stability of Commonly Evaluated Biomarkers in Clinically Stable COPD. *American Journal of Respiratory and Critical Care Medicine Conference: International Conference of the American Thoracic Society, ATS*. 2022;205(1).
564. Platz EA, Sutcliffe S, De Marzo AM, Drake CG, Rifai N, Hsing AW, et al. Intra-individual variation in serum C-reactive protein over 4 years: an implication for epidemiologic studies. *Cancer Causes & Control*. 2010;21(6):847-51.
565. Qi Z, Chen Y, Zhang L, Ma X, Wang F, Cheng Q, et al. Biological variations of thirteen plasma biochemical indicators. *Clinica Chimica Acta*. 2016;452:87-91.
566. Riese H, Vrijkotte TG, Meijer P, Kluft C, de Geus EJ. Diagnostic strategies for C-reactive protein. *BMC Cardiovascular Disorders*. 2002;2:9.
567. Rubin LH, Benning L, Keating SM, Norris PJ, Burke-Miller J, Savarese A, et al. Variability in C-reactive protein is associated with cognitive impairment in women living with and without HIV: a longitudinal study. *Journal of Neurovirology*. 2018;24(1):41-51.
568. Rudez G, Meijer P, Spronk HM, Leebeek FW, ten Cate H, Kluft C, et al. Biological variation in inflammatory and hemostatic markers. *Journal of Thrombosis & Haemostasis*. 2009;7(8):1247-55.

569. Sakkinen PA, Macy EM, Callas PW, Cornell ES, Hayes TE, Kuller LH, et al. Analytical and biologic variability in measures of hemostasis, fibrinolysis, and inflammation: assessment and implications for epidemiology. *American Journal of Epidemiology*. 1999;149(3):261-7.
570. Sennels HP, Jacobsen S, Jensen T, Hansen MS, Ostergaard M, Nielsen HJ, et al. Biological variation and reference intervals for circulating osteopontin, osteoprotegerin, total soluble receptor activator of nuclear factor kappa B ligand and high-sensitivity C-reactive protein. *Scandinavian Journal of Clinical & Laboratory Investigation*. 2007;67(8):821-35.
571. Sjoberg B, Snaedal S, Stenvinkel P, Qureshi AR, Heimbürger O, Barany P. Three-month variation of plasma pentraxin 3 compared with C-reactive protein, albumin and homocysteine levels in haemodialysis patients. *Clinical Kidney Journal*. 2014;7(4):373-9.
572. Thompson AE, Pope JE. The erratic C-reactive protein: a novel outcome measure for longitudinal disease activity in rheumatoid arthritis. *Clinical and Experimental Rheumatology*. 2022;40(7):1411-6.
573. Todd J, Simpson P, Estis J, Torres V, Wub AH. Reference range and short- and long-term biological variation of interleukin (IL)-6, IL-17A and tissue necrosis factor-alpha using high sensitivity assays. *Cytokine*. 2013;64(3):660-5.
574. Tsirpanlis G, Bagos P, Ioannou D, Bleta A, Marinou I, Lagouranis A, et al. The variability and accurate assessment of microinflammation in haemodialysis patients. *Nephrology Dialysis Transplantation*. 2004;19(1):150-7.
575. van den Berg VJ, Umans VAWM, Brankovic M, Oemrawsingh RM, Asselbergs FW, van der Harst P, et al. Stabilization patterns and variability of hs-CRP, NT-proBNP and ST2 during 1 year after acute coronary syndrome admission: results of the BIOMarCS study. *Clinical chemistry and laboratory medicine*. 2020;08.
576. Waschki B, Schaper M, Sack A, Paasch K, Mueller K, Feindt B, et al. Cardiovascular biomarkers in COPD patients during three years of follow-up. *American Journal of Respiratory and Critical Care Medicine Conference: American Thoracic Society International Conference, ATS*. 2010;181(1 MeetingAbstracts).
577. Wu S, Li Y, Jin C, Yang P, Li D, Li H, et al. Intra-individual variability of high-sensitivity C-reactive protein in Chinese general population. *International Journal of Cardiology*. 2012;157(1):75-9.
578. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care*. 2004;42(12):1194-201.
579. Yeung AS, Jing Y, Brennenman SK, Chang TE, Baer L, Hebden T, et al. CLINICAL OUTCOMES IN MEASUREMENT-BASED TREATMENT (COMET): A TRIAL OF DEPRESSION MONITORING AND FEEDBACK TO PRIMARY CARE PHYSICIANS. *Depress Anxiety*. 2012;29(10):865-73.
580. Moore M, Ali S, Stuart B, Leydon GM, Ovens J, Goodall C, et al. Depression management in primary care: an observational study of management changes related to PHQ-9 score for depression monitoring. *Br J Gen Pract*. 2012;62(599):e451-e7.
581. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ*. 2011;184(3):E191-E6.
582. Kroenke K, Spitzer R, L. , Williams J, B. W. The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener. *Med Care*. 2003;41(11):1284-92.
583. Bocéréan C, Dupret E. A validation study of the Hospital Anxiety and Depression Scale (HADS) in a large sample of French employees. *BMC Psychiatry*. 2014;14(1):354-.
584. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983;67(6):361-70.
585. Wilkinson MJ, Barczak P. Psychiatric screening in general practice: comparison of the general health questionnaire and the hospital anxiety depression scale. *J R Coll Gen Pract*. 1988;38(312):311-3.

586. Dozois DJA, Dobson KS, Ahnberg JL. A Psychometric Evaluation of the Beck Depression Inventory-II. *Psychological Assessment*. 1998;10(2):83-9.
587. von Glischinski M. How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). *Qual Life Res*. 28(5):1111-9.
588. Bakalidou D, Skordilis EK, Giannopoulos S, Stamboulis E, Voumvourakis K. Validity and reliability of the FSS in Greek MS patients. *Springerplus*. 2013;2(1):304.
589. Boyer FC, Rapin A, Calmus A, Percebois-Macadre L, Tambosco L, Bertaud S, et al. HADS scale in adults suffering from Steinert myotonia: Reproducibility and internal consistency. *Annals of Physical and Rehabilitation Medicine*. 2011;54:e238.
590. Chen IP, Liu SI, Huang HC, Sun FJ, Huang CR, Sung MR, et al. Validation of the Patient Health Questionnaire for Depression Screening Among the Elderly Patients in Taiwan. *International Journal of Gerontology*. 2016;10(4):193-7.
591. Chiu EC, Chen YJ, Wu WC, Chou CX, Yu MY. Psychometric Comparisons of Three Depression Measures for Patients With Stroke. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association*. 2022;76(4).
592. Ganguly S, Samanta M, Roy P, Chatterjee S, Kaplan DW, Basu B. Patient health questionnaire-9 as an effective tool for screening of depression among indian adolescents. *Journal of Adolescent Health*. 2013;52(5):546-51.
593. Gelaye B. Diagnostic validity of patient health questionnaire-9 and composite international diagnostic interview assessment scales for depression in East Africa. *Dissertation Abstracts International: Section B: The Sciences and Engineering*. 2014;74(9-B(E)):No Pagination Specified.
594. Hansen H, Bieler T, Beyer N, Frolich A, Kallemose T, Godtfedsen N. How does EQ-5D-3L and HADS questionnaires perform in patients with COPD? *European Respiratory Journal Conference: European Respiratory Society International Congress, ERS*. 2020;56(Supplement 64).
595. Hitchon CA, Zhang L, Peschken CA, Lix LM, Graff LA, Fisk JD, et al. The validity and reliability of screening measures for depression and anxiety disorders in rheumatoid arthritis. *Arthritis care & research*. 2019;14.
596. Huang SL, Hsieh CL, Wu RM, Lu WS. Test-retest reliability and minimal detectable change of the Beck Depression Inventory and the Taiwan Geriatric Depression Scale in patients with Parkinson's disease. *PLoS ONE*. 2017;12(9).
597. Jin H, Wu S. Text Messaging as a Screening Tool for Depression and Related Conditions in Underserved, Predominantly Minority Safety Net Primary Care Patients: Validity Study. *Journal of medical Internet research*. 2020;22(3):e17282.
598. Kha TV, Stenager E, Hoang H, Bruun-Plesner K, Fuglsang KS, la Cour BS, et al. Preliminary validity and test-retest reliability of two depression questionnaires compared with a diagnostic interview in 99 patients with chronic pain seeking specialist pain treatment. *Scandinavian Journal of Pain*. 2020;20(4):717-26.
599. Lamers F, Jonkers CCM, Bosma H, Penninx BWJH, Knottnerus JA, van Eijk JTM. Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. *Journal of Clinical Epidemiology*. 2008;61(7):679-87.
600. Li W, Kai L, Jianchao L, Li S, Rongjing D, Dayi H. Value of patient health questionnaires (PHQ)-9 and PHQ-2 for screening depression disorders in cardiovascular outpatients. [Chinese]. *Zhonghua xin xue guan bing za zhi*. 2015;43(5):428-31.
601. Lodhi FS, Elsous AM, Irum S, Khan AA, Rabbani U. Psychometric properties of the Urdu version of the Hospital Anxiety and Depression Scale (HADS) among pregnant women in Abbottabad, Pakistan. *Gen Psychiatr*. 2020;33(5):e100276.

602. Marinus J, Leentjens AFG, Visser M, Stiggelbout AM, Van Hilten JJ. Evaluation of the hospital anxiety and depression scale in patients with Parkinson's disease. *Clinical Neuropharmacology*. 2002;25(6):318-24.
603. Michopoulos I, Douzenis A, Kalkavoura C, Christodoulou C, Michalopoulou P, Kalemi G, et al. Hospital anxiety and depression scale (HADS): Validation in a Greek general hospital sample. *Annals of General Psychiatry*. 2008;7 (no pagination).
604. Mohamadian R, Khazaie H, Ahmadi SM, Fatmizade M, Ghahremani S, Sadeghi H, et al. The psychometric properties of the persian versions of the patient health questionnaires 9 and 2 as screening tools for detecting depression among university students. *International Journal of Preventive Medicine*. 2022;13(1):116.
605. Monahan PO, Shacham E, Reece M, Kroenke K, Ong'Or WO, Omollo O, et al. Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in Western Kenya. *Journal of General Internal Medicine*. 2009;24(2):189-97.
606. Omoro S, Fann J, Weymuller E, Macharia I, Yueh B. Swahili translation and validation of the Patient Health Questionnaire-9 depression scale in the Kenyan head and neck cancer patient population. *International Journal of Psychiatry in Medicine*. 2006;36(3):367-81.
607. Pinto-Meza A, Serrano-Blanco A, Penarrubia MT, Blanco E, Haro JM. Assessing depression in primary care with the PHQ-9: Can it be carried out over the telephone? *Journal of General Internal Medicine*. 2005;20(8):738-42.
608. Rae CS, Tsangaris E, Klassen AF, Breakey V, D'Agostino N. Comparison of Patient-Reported Outcome Measures for Use as Performance Metrics in Adolescent and Young Adult Psychosocial Cancer Care. *Journal of Adolescent and Young Adult Oncology*. 2020;9(2):262-70.
609. Reda AA. Reliability and validity of the Ethiopian version of the hospital anxiety and depression scale (HADS) in HIV infected patients. *PLoS ONE*. 2011;6(1).
610. Silverberg JI, Lee B, Lei D, Yousaf M, Janmohamed SR, Vakharia PP, et al. Measurement Properties of Patient Health Questionnaire 9 and Patient Health Questionnaire 2 in Adult Patients with Atopic Dermatitis. *Dermatitis*. 2021;32(4):225-31.
611. Summaka M, Zein H, Abbas LA, Elias C, Elias E, Fares Y, et al. Validity and Reliability of the Arabic Patient Health Questionnaire-9 in Patients with Spinal Cord Injury in Lebanon. *World Neurosurgery*. 2019;125:e1016-e22.
612. Tanveer Hasan ATM, Haq SA, Ahmed S, Abdal SJ, Majumder MSM, Nahiduzzamane Shazzad M, et al. Translation & cross-cultural adaptation of the english patient health questionnaire-9 (PHQ-9) to bangla and its validation in adult SLE patients. *International Journal of Rheumatic Diseases*. 2017;20 (Supplement 1):57.
613. Visser M, Leentjens AFG, Marinus J, Stiggelbout AM, van Hilten JJ. Reliability and validity of the Beck Depression Inventory in patients with Parkinson's disease. *Movement Disorders*. 2006;21(5):668-72.
614. Wang W, Chair SY, Thompson DR, Twinn SF. A psychometric evaluation of the Chinese version of the Hospital Anxiety and Depression Scale in patients with coronary heart disease. *J Clin Nurs*. 2009;18(13):1908-15.
615. Woldetensay YK, Belachew T, Tesfaye M, Spielman K, Biesalski HK, Kantelhardt EJ, et al. Validation of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression in pregnant women: Afaan Oromo version. *PLoS ONE*. 2018;13(2).
616. Yu X, Stewart SM, Wong PT, Lam TH. Screening for depression with the Patient Health Questionnaire-2 (PHQ-2) among the general population in Hong Kong. *Journal of Affective Disorders*. 2011;134(1-3):444-7.
617. Yu X, Tam WW, Wong PT, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. *Compr Psychiatry*. 2012;53(1):95-102.