STATISTICAL METHODS FOR ASSESSING THE REPRODUCIBILITY OF BIOMARKERS

by

KONSTANTINOS TRYPOSKIADIS

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
September 2022

Abstract

Introduction

Biomarkers are often subject to measurement error, affecting their reproducibility. Statistical methods for examining the reproducibility of continuous biomarkers are well-developed. However, methods for performing systematic reviews and meta-analyses of biomarker reproducibility, and examining the reproducibility of count-based biomarkers are under-researched.

Aims

To propose methods for performing meta-analyses of reproducibility of continuous measurements, and for examining the reproducibility of count measurements.

Methods

Current methods for performing systematic reviews and meta-analyses of reproducibility of continuous measurements were systematically identified and critiqued. Meta-analytic methods were developed and evaluated in a case study examining the reproducibility of grip strength measured in different populations. Methods for count outcomes were evaluated in a case study of a biopsy-based biomarker for Sjorgren's syndrome. Simulation extended the case study to examine how reproducibility varied for larger numbers of biopsied samples.

Results

Methods for meta-analysis performed well, indicating that grip strength measurements are reproducible. Count-based methods outperformed continuous-based methods when

examining the reproducibility of biopsy-based count measurements. The methods indicated low reproducibility, with simulation suggesting improved reproducibility for larger samples.

Conclusion

The proposed methods provide robust summary and primary evidence of the reproducibility of continuous and count-based biomarkers, respectively, allowing better decisions regarding their use in practice.

Acknowledgements

This PhD was funded by the National Institute of Health Research (NIHR) Birmingham Biomedical Research Centre (BRC). I am extremely grateful to my supervisors, Professor Jon Deeks, Dr Alice Sitch, Dr Jac Dinnes, and Dr Malcolm Price, for giving me the opportunity to work on this wonderful project, the interesting discussions, the supervision, and comments on my chapters.

I am indebted to Jac Dinnes, April Coombe, Laura Quinn, Sayeed Haque, and Alex Gough for their significant contributions to the review work.

I am also grateful to Dr Ben Fisher and Dr Amrit Dhaliwal for providing the data sets used as case studies in the thesis, and their clinical advice.

Finally, a special thanks to my wonderful parents, Filippos and Katerina, brother and sister, Thodoris and Polina, partner, Danai, and friends, for the support, encouragement, and interest on my wellbeing during these challenging years.

Contents

1. Introduction 1
1.1. Definition and applications of biomarkers 1
1.2. Requirements for the use of biomarkers in medical research and practice 2
1.3. Reproducibility and measurement error of biomarkers 3
1.3.1. The measurement error 3
1.3.2. The impact of measurement error on medical research and practice 4
1.3.3. Potential sources of test variability examined in primary studies 5
1.3.4. Issues in the field of reliability and measurement error of biomarkers 9
1.4. Aim and objectives of the thesis 10
1.5. Thesis outline 10
2. Background on primary studies examining the measurement error of biomarkers, and the meta-analysis of estimates reported in primary studies 13
2.1. Introduction 13
2.2. Design of primary studies examining the measurement error of biomarkers 13
2.2.1. Minimising pre-analytical variability prior to testing 17
2.2.2. Populations investigated 17
2.2.3. Time interval between measurements 18
2.2.4. Sample size 19
2.3. Statistical analysis of primary studies examining the reliability and measurement error of biomarkers 21
2.3.1. Preparation of the data prior to analysis 22
2.3.2. General method used for estimating sources of variability in the measurements 28
2.3.3. Statistical parameters reported in primary studies examining measurement error 35
2.3.4. Additional methods for constructing confidence intervals 57
2.4. Reporting of primary studies examining the measurement error of biomarkers 61
2.5. General statistical methods used for the meta-analysis of data reported across primary studies 64
2.5.1. Meta-analysis using aggregate study-level data 64
2.5.2. Meta-analysis using individual participant data 70
2.6. Discussion 71
2.7. Conclusion 74
3. Evaluation of the reproducibility of grip strength measurements produced by hand-held digital dynamometers. 75
3.1. Introduction 75

3.2. Clinical background 75
3.3. Study design 76
3.4. Components of variability in the measurements 77
3.5. Aim and objectives 78
3.6. Statistical methods 79
3.6.1. Description of statistical model used to estimate the components of variability 80
3.6.2. Regression-based parameters of reliability and measurement error 81
3.6.3. Additional parameters used for pairwise comparisons 85
3.7. Results 88
3.7.1. Regression-based parameters of reliability and measurement error 92
3.7.2. Additional parameters used for pairwise comparisons 95
3.8. Discussion 101
3.9. Conclusion 104
4. Current state of systematic review methods and meta-analytic approaches used for evaluating the reliability and measurement error of biomarkers 105
4.1. Introduction 105
4.2. Objectives 106
4.3. Methods 107
4.4. Results 109
4.4.1. Summary of Reviews identified 109
4.4.2. Description of review methods 111
4.4.3. Description of statistical estimates reported in the reviews 113
4.4.4. Description of pooling approaches used in the reviews 114
4.5. Discussion 127
4.6. Conclusion 131
5. Proposed meta-analytic approaches for parameters of measurement error 133
5.1. Introduction 133
5.2. Aim 134
5.3. Proposed meta-analytic approach for the limits of agreement 134
5.4. Proposed meta-analytic approach for the standard error of measurement 137
5.5. Proposed meta-analytic approach for the coefficient of variation 140
5.6. Discussion 142
5.7. Conclusion 143
6. Systematic Review and meta-analysis of the reliability and measurement error of hand-held dynamometers used to assess grip strength 144
6.1. Introduction 144

6.2. Objectives 144
6.3. Review methods 145
6.3.1. Inclusion criteria 145
6.3.2. Search strategy 146
6.3.3. Selection of studies 147
6.3.4. Data extraction 147
6.3.5. Assessment of methodological quality 148
6.4. Decisions made when multiple estimates are reported within studies 149
6.5. Statistical methods 152
6.5.1. Statistical model used for the meta-analysis of the reported estimates 152
6.5.2. Approach used for the intra class correlation coefficient 153
6.5.3. Approach used for the Pearson correlation coefficient 154
6.5.4. Approach used for the standard error of measurement 154
6.5.5. Approach used for the smallest detectable change 155
6.5.6. Approach used for the limits of agreement 155
6.5.7. Approach used for the coefficient of variation 157
6.5.8. Subgroup analysis 158
6.6. Results 159
6.6.1. Summary of identified studies 159
6.6.2. Assessment of methodological quality 164
6.6.3. Reproducibility of measurements taken at different testing sessions 165
6.6.4. Reproducibility of measurements taken within the same testing session 180
6.6.5. Subgroup analysis 182
6.7. Discussion 184
6.8. Conclusion 188
7. Alternative statistical methods for estimating sources of variability in the measurements of count-based biomarkers 189
7.1. Introduction 189
7.2. Aim 189
7.3. Statistical models used for analysing multi-level count data 190
7.3.1. The Poisson distribution 190
7.3.2. The negative binomial distribution 193
7.3.3. Which model should be used? 196
7.4. Parameters of reliability and measurement error 197
7.4.1. The standard error of measurement (SEM) 199
7.4.2. The coefficient of variation (CV) 200

7.4.3. The intra class correlation (ICC) 200
7.4.4. The median rate ratio (MRR) 201
7.5. Discussion 202
7.6. Conclusion 203
8. Application of statistical methods appropriate for estimating sources of variability in count-based biomarkers 204
8.1. Introduction 204
8.2. Clinical background 204
8.3. Components of variability in the focus score 205
8.4. Objectives 206
8.5. Statistical methods 206
8.5.1. Analysis using a two-level Poisson model 207
8.5.2. Analysis using a two-level negative binomial model 208
8.5.3. Analysis using a two-level linear model 209
8.5.4. Method for comparing the fitted models 210
8.5.5. Parameters of reliability and measurement error 211
8.6. Results 212
8.6.1. Results from two-level Poisson model 216
8.6.2. Results from two-level negative binomial model 217
8.6.3. Results from two-level linear model 217
8.6.4. Model comparison 218
8.7. Discussion 219
8.8. Conclusion 220
9. Evaluation of the impact of the number of glands within labial salivary gland biopsy on the reproducibility of the focus score 222
9.1. Introduction 222
9.2. Aim 223
9.3. Statistical methods 223
9.3.1. Simulation models 223
9.3.2. Simulation inputs 226
9.3.3. Analysis of generated data 230
9.3.4. Justification of the number of simulations 231
9.4. Results 232
9.4.1 Base-case scenario 232
9.4.2. Sensitivity analysis 239
9.5. Discussion 250

9.6. Conclusion 251
10. Discussion and conclusions 252
10.1. Thesis overview and summary of findings 252
10.2. Strengths and limitations 257
10.2.1. Strengths 257
10.2.2. Limitations 259
10.3. Implications for medical research and practice 262
10.4. Further work 264
10.5. Conclusions 265
References 267
Appendices 293
Appendix A: Additional figures for Chapter 3 293
Appendix A1. Histogram of the difference between the mean values produced at the first and second visit. 293
Appendix A2. Histogram of the difference between the highest values produced at the first and second visit. 294
Appendix A3. Scatterplot of the mean of 3 within-individual measurements produced at each visit. 295
Appendix A4. Scatterplot of the highest of 3 within-individual measurements produced at each visit. 296
Appendix B: Search strategy and additional tables for Chapter 4 297
Appendix B1. Search strategy. 297
Appendix B2. Data extraction form. 299
Appendix B3. Characteristics of the identified systematic reviews. 300
Appendix B4. Details on quality items for each systematic review. 347
Appendix C: Search strategy and additional tables for Chapter 6 370
Appendix C1. Search strategy. 370
Appendix C2. Data extraction form. 373
Appendix C3. Standards on design requirements and statistical methods for studies on reliability or measurement error. After Mokkink et al [37]. 374
Appendix C4. Data extraction - Characteristics of the identified primary studies. 376
Appendix C5. Data extraction (continued) - Characteristics of study individuals and participating observers. 384
Appendix C6. Data extraction (continued) - Measurement conditions and protocol used. 391
Appendix C7. Quality assessment of the identified primary studies. 403

List of Tables

Table 1.1. Terms used in the thesis. 6
Table 2.1. ANOVA-based calculations for estimating the components of variability in laboratory tests 29
Table 2.2. Methods for estimating the components of variability in non-laboratory tests 33
Table 2.3. Commonly reported parameters of reliability and measurement error. 35
Table 2.4. Interpretation of the intra class correlation based on Cicchetti [52], and Koo and Li [53]. 42
Table 2.5. Methods for calculating 95% confidence intervals for all different types of intra class correlations. 43
Table 2.6. Interpretation of the Kappa value based on Landis and Koch [71], and Fleiss [72]. 54
Table 2.7. Example of two observers assessing presence or absent of a condition. 55
Table 2.8. Reporting guidelines provided by Bartlett et al [85] and Kottner et al [36]. 62
Table 2.9. Interpretation of I-squared value based on the Cochrane Handbook. 67
Table 3.1. Baseline characteristics of 84 patients recruited in the study. 90
Table 3.2. Results obtained from linear regression analyses performed for each hand. 94
Table 3.3. Results obtained from pairwise comparisons performed between the two visits. 100
Table 3.4. Results obtained from pairwise comparisons performed within each visit. 100
Table 4.1. Summary of the identified reviews. 111
Table 4.2. Assessment of the quality of the review process used in the identified reviews. 112
Table 4.3. Statistical parameters reported in the identified reviews. 113
Table 4.4. Summary of synthesis methods used in the identified reviews. 126
Table 6.1. Summary of the identified studies (N=80). 161
Table 6.2. Primary analysis including studies examining the intra-observer reproducibility across 2 testing sessions. 177
Table 6.3. Primary analysis including studies examining the inter-observer reproducibility across 2 testing sessions. 179
Table 6.4. Primary analysis of studies examining the reproducibility within testing sessions. 181
Table 6.5. Subgroup analysis of studies examining the reproducibility across two testing sessions. 183
Table 7.1. Estimation of parameters for each statistical model presented. 198
Table 8.1. Estimation of the grand mean and variance components across the three models. 210
Table 8.2. Characteristics of patients recruited in the OASIS study (N=32). 215
Table 8.3. Results obtained from the analysis of the OASIS cohort (N=32). 218
Table 9.1. Notation description and estimation method for simulation parameters. 229
Table 9.2. Generation of gland-level parameters – base case scenario. 233
Table 9.3. Reproducibility of the focus score – base case scenario. 233
Table 9.4. Generation of gland-level parameters – varying the sample size. 240
Table 9.5. Reproducibility of the focus score – varying the sample size. 241
Table 9.6. Generation of gland-level parameters – varying the regression parameters required for the number of foci within glands. 242
Table 9.7. Reproducibility of the focus score – varying the regression parameters required for the number of foci within glands. 244

Table 9.8. Generation of gland-level parameters – varying the regression parameters required for the square root of the area of the glands. 246
Table 9.9. Reproducibility of the focus score – varying the regression parameters required for the square root of the area of the glands. 248

List of Figures

Figure 2.1. Design of biological variability study with a total of n_G participants and n_I measurements produced from each participant. 15
Figure 2.2. Impact of the grand mean, the analytical and within-individual biological standard deviations on the number of samples required for a mean of multiple within-individual measurements lying within $D\%$ of the true value of the individual. 20
Figure 2.3. Impact of the number of within-individual measurements and the pre-specified ICC and 95% confidence interval on the number of individuals required. 21
Figure 2.4. Detection of outliers as described in Braga and Panteghini [27]. 24
Figure 2.5. Normality checking as described in Braga and Panteghini [27]. 26
Figure 2.6. Two examples of Bland-Altman plots of the difference between two measurements produced from the same individual against the corresponding average value. 27
Figure 3.1. Study design. 79
Figure 3.2. Flow chart of patients, showing the numbers who provided data at each visit. 89
Figure 3.3. Distribution of the mean of multiple measurements produced from each patient, by hand. 91
Figure 3.4. Bland Altman plots of the mean and highest of three within-individual measurements produced at each visit. 98
Figure 4.1. PRISMA flow chart. 110
Figure 6.1. PRISMA flow chart. 161
Figure 6.2. COSMIN Risk of Bias tool – Standards for study design. 164
Figure 6.3. Distribution of reported intra class correlations and the corresponding Fisher's Z-values. 166
Figure 6.4. Forest plot of Fisher's Z (for intra class correlations) including studies examining intra-observer reproducibility across two different testing sessions. 167
Figure 6.5. Forest plot of Fisher's Z (for Pearson correlations) including studies examining intra-observer reproducibility across two different testing sessions. 168
Figure 6.6. Distribution of values reported for the standard error of measurement (SEM) and corresponding logarithm of squared values ($\log(SEM^2)$). 170
Figure 6.7. Forest plot of $\log SEM^2$ including studies examining intra-observer reproducibility across two different testing sessions. 171
Figure 6.8. Distribution of values reported for the standard deviation of difference (SD_d) and corresponding logarithm of squared values ($\log(SD_d)^2$) 173
Figure 6.9. Forest plot of the two parameters required for the limits of agreement (d and $\log(SD_d)^2$) including studies examining intra-observer reproducibility across two different testing sessions. 174
Figure 6.10. Forest plot of the two parameters required for the coefficient of variation ($\log(SEM^2)$ and μ) including studies examining intra-observer reproducibility across two different testing sessions. 176
Figure 8.1. Microphotograph illustrating LSG biopsy obtained from a patient with confirmed disease. The total number of foci (black outlined area) is 8. The measured glandular area (in red) is 20.89mm^2. This gives a focus score of 1.53 for the patient. Graph taken from Fisher et al [210]. 205
Figure 8.2. Distribution of the number of foci observed within glands and the area of the glands. 214

Figure 8.3. Dot plot of the observed focus score of each gland, split by patient. A blue dot represents the focus score of each gland. A red cross represents the median value within patient. The red line represents the mean focus score (=1.79). 216
Figure 9.1. Distribution of the produced 2500 estimates for the mean of the number of foci within glands and each LSG biopsy – base case scenario. 234
Figure 9.2. Distribution of the produced 2500 estimates for the mean of the square root of the area of the glands and each LSG biopsy – base case scenario. 235
Figure 9.3. Scatterplot of the 2500 median differences in the focus score against the associated interquartile range – base case scenario. The reference value of median=0.5 is used to aid comparissons between different simulated numbers of glands. 236
Figure 9.4. Boxplot of median absolute difference in focus score for each simulated number of glands – base case scenario. The reference value of median=0.5 is used to aid comparissons between different simulated numbers of glands. 237
Figure 9.5. Boxplot of IQR of absolute difference in focus score for each simulated number of glands – base case scenario. The reference value of IQR=1 is used to aid comparissons between different simulated numbers of glands. 238

1. Introduction

Biomarkers indicate the underlying medical state of an individual, and are often used to evaluate the presence/progress of a medical condition, the effects and safety of new interventions, or the occurrence of future clinical outcomes [1, 2]. In order to be used in medical research and practice, it is essential that biomarkers provide reproducible measurements.

The broad aim of this thesis is to propose statistical methods for evaluating the reproducibility of biomarkers, before biomarkers are used in medical research and practice. The thesis mainly focuses on i) bringing together estimates of the reproducibility of biomarkers from multiple primary studies (meta-analysis); and ii) the statistical analysis of primary studies examining the reproducibility of biomarkers expressed as counts rather than continuous measurements.

1.1. Definition and applications of biomarkers

Biomarkers (short for biological markers) are defined as measurements used to evaluate potential chemical, physical or biological hazards, normal biological processes, and the effectiveness of new therapeutic interventions [3, 4]. Applications of biomarkers in medical research and practice include screening for disease; the diagnosis, prognosis, and monitoring of disease; the stratification of patients according to disease severity; and evaluation of the effects and safety of new interventions or environmental agents. Furthermore, the use of biomarkers as surrogate endpoints (i.e., substitutes of clinically meaningful outcomes) in clinical trials has become commonplace, and has been approved by the Food and Drug Administration (FDA) [1, 5]. The measured response may be produced from molecular, histologic, imaging, or physiologic tests [5]. Examples of the use of biomarkers in clinical practice include:

- blood pressure readings as a diagnostic biomarker of hypertension [6].
- faecal occult blood test as a screening biomarker for colorectal cancer [7].

- prostate-specific antigen (PSA) as a prognostic biomarker for prostate cancer [8].
- cancer antigen 125 (CA 125) as a monitoring biomarker for assessing disease status or burden during and after treatment in patients with ovarian cancer [9].
- left ventricular ejection fraction (LVEF) as a stratification biomarker for the type of heart failure [10].
- serum potassium as a biomarker for evaluating the safety of patients on diuretics [11].
- serum creatinine as a surrogate endpoint in a clinical trial of patients with atherosclerotic renovascular disease (ARVD), evaluating whether revascularisation (with angioplasty and/or stent) can prevent or delay the progression from ARVD to ESRD (i.e., end-stage renal disease), compared to the standard care [12].

1.2. Requirements for the use of biomarkers in medical research and practice

When considered for use in medical research and practice, the suitability of biomarkers is often taken for granted [1]. However, strong scientific evidence is required, as not all biomarkers are intended for this purpose. Strimbu and Tavel [1], and Califf [2] discuss several criteria that need to be met so that biomarkers qualify for use in medical research and practice, which should be examined simultaneously. Strimbu and Tavel highlight the importance of relevance and validity [1]. Relevance refers to the ability of biomarkers to provide clinically relevant information on questions that are of interest to the health care providers, health policy makers, and the public. Validity refers to the ability of biomarkers to measure what they are intended to (i.e., a strong correlation with the clinical endpoint of interest should be observed). Califf additionally states that biomarkers should not only correlate with the clinical outcome of interest, but also be able to detect changes over time [2]. This means that a change in the biomarker should also correspond to a change in the clinical outcome. Another essential requirement for use in medical research and practice is the reproducibility of biomarkers, defined as the extent to which two or more measurements produced for the same

individual are the same, given that the health status of the individual has not changed in between the measurements [13].

1.3. Reproducibility and measurement error of biomarkers

The evaluation of the reproducibility of biomarkers prior to being used in medical research and practice is the main theme of the thesis. When biomarkers are measured two or more consecutive times in individuals with stable health status, the produced measurements are expected to be very similar, if not identical. However, this is not often the case, as each measurement is often subject to measurement error and may not accurately reflect the underlying true disease state [13].

1.3.1. The measurement error

In theory, measurement error is defined as the absolute difference between an observed measurement produced for an individual, and the true value of the individual [14, 15]. However, as the true value is nearly always unknown, common practice often involves taking multiple measurements from an individual, and using the mean of the measurements as the best estimate of the true value [15]. The measurement error in turn reflects the variability of the produced measurements around the mean value of the individual. Measurement error may arise due to:

- **Systematic variability occurring between measurements.** That is, a general trend for repeated measurements to be different in a particular (positive or negative) direction [16]. Systematic variability may for example arise due to inconsistencies in the measurement protocol (e.g., the posture of an individual not being consistent across different testing occasions), two clinicians assessing the same individual in a different way (e.g., for imaging tests, two clinicians may rate the same image differently), or even true differences occurring within the patients over time (e.g., patients performing better at the second testing occasion

due to their experience with the first). When examining the measurement error of biomarkers, every possible effort should be made so that any potential systematic variability between measurements is eliminated. However, in some cases, potential systematic variability between two different clinicians (or even within the same clinician) may be of interest (see section 1.3.3).

- **Random variability occurring between measurements.** Even if all potential sources of systematic variability are eliminated, two consecutive measurements may still differ due to the random error occurring within each measurement, which may be subject to factors that are often not possible to be controlled [16]. Examples include patients feeling more relaxed when performing the second measurement compared to the initial, unexpected changes in the temperature or the environment, or the inherent variability that a measurement tool may have [16, 17].

1.3.2. The impact of measurement error on medical research and practice

Using biomarkers of low reproducibility (i.e., high measurement error) may lead to false conclusions with respect to the diagnosis or the classification of a medical condition, the effects and safety of treatments, or the occurrence of future clinical outcomes. A few examples of the potential impact of measurement error on medical research and practice include:

- The measurement of blood pressure, which is commonly used for the diagnosis of hypertension [6], is considered a key prognostic factor in the development of cardiovascular risk scores [18, 19], and has also been used as an outcome in clinical trials [12]. However, measurements of blood pressure in clinical practice are known to be subject to multiple sources of error (e.g., incorrect positioning of individuals during the assessment, inadequate equipment, improper cuff bladder size, incorrect technique being used, or even random

within-individual variability due to biological factors) [15, 20], which may in turn lead to the over/under treatment of patients [20].

- The ultrasound-based measurement of the cross-sectional area (CSA) of peripheral nerves, which is validated for the diagnosis of Carpal Tunnel Syndrome [21]. Gao et al examined the percentage error from a known measurement among nine ultrasound examiners, which was found to be approximately 10% [21]. Given that a cut-off value of $10mm^2$ is commonly used to define presence or absence of the condition, the authors state that measurements lying within 10% of the diagnostic threshold may lead to the misdiagnosis of the condition, and should be interpreted with caution [21].
- The measurement of the left ventricular ejection fraction (LVEF) based on echocardiography, which is used as a guide for the current management of patients with chronic heart failure [22, 23]. A value of ≤40% has been used to define patients with a reduced ejection fraction [24, 25], with guidelines recommending the use of combination treatment with neurohormonal antagonists for such patients [22]. However, evidence suggests a variability up to 15% when experts in echocardiography read the same image [22], which may lead to patients being misclassified, and in turn treated inappropriately [22, 23].

1.3.3. Potential sources of test variability examined in primary studies

When biomarkers are considered for use in medical research and practice, researchers and medical professionals need to be familiar with the potential sources of variability that is inherent to the produced measurements (i.e., the measurement error). As such, primary studies are often designed to examine any such sources, and estimate the degree of error attributed to each source (see Chapter 2 for a detailed description of the study design and statistical analysis). The potential sources of the inherent variability in the measurements may differ depending on the type of test being used. See Table 1.1 for a guide to the terminology used in this thesis.

Table 1.1. Terms used in the thesis.

Reproducibility	The extent to which two or more measurements produced within individuals are the same, given that the health status of the individuals has not changed in between the measurements.
Measurement error	The variability of multiple measurements produced from the same individual around the mean of the measurements (which serves as the best estimate of the true value of the individual).
Reliability	The ability of a test to distinguish patients from each other despite the presence of measurement error.
Random variability between measurements	The variability in measurements produced within individuals due to any random analytical and/or within-individual biological variability.
- Analytical variability	The inherent variability of the equipment used for producing the measurements.
- Within-individual biological variability	The random fluctuations around the true value of an individual.
Systematic variability between measurements	A general trend for measurements produced within individuals to be different in a particular (positive or negative) direction.
- Pre-analytical variability	The variability in measurements produced within individuals due to any incomplete preparatory actions required prior to taking a measurement, or due to how measurements have been obtained and handled.
- Inter-observer variability	The variability in measurements produced within individuals due to systematic differences observed between two or more observers assessing the same individual.

- Intra-observer variability	The variability in measurements produced within individuals due to systematic differences observed within the same observer assessing the same individual multiple times.
Between-individual biological variability	The variability in the true value of different individuals.

Studies of laboratory tests

In 1989, Fraser and Harris provided a framework for the design and analysis of studies examining measurement error conducted in the clinical chemistry laboratory setting [26]. The Fraser-Harris framework was updated by Braga and Panteghini in 2016 [27]. The authors focus on four potential sources of variability when considering laboratory-based tests [26, 27]. These include:

- **Pre-analytical variability,** which refers to systematic variability arising due to the preparation of the patients prior to the collection of a sample (fasting, starvation, exercise, altitude, incorrect posture during sample collection), as well as the sample collection and handling (e.g., prolonged tourniquet application, transportation time, centrifugation time, inappropriate storage conditions prior to analysis) [26, 27].
- **Analytical variability,** which refers to the variability observed when the analysis of the same sample is replicated, and includes both random and systematic analytical variability. Random analytical variability refers to the inherent variability that every analytical technique has; whereas systematic analytical variability occurs when major changes to in the instrumentation or the methodology are made during the study. Systematic analytical variability may be caused by changes in calibration lots, reagent lots, or operators (as one may perform tasks consistently but differently compared to others) [26, 27].

- **Within-individual biological variability**, which refers to any random fluctuations around the homeostatic setting point of the same individual. The homeostatic setting point refers to the true value of the biomarker for each individual. These fluctuations may occur at different times of the day (e.g., changes in the sleep/wake cycle immediately affect growth hormone concentrations), different times during the month (e.g., the breast tumour biomarker CA-153 has monthly cycles), or even at different seasons (e.g., blood volume increases with higher temperatures) [26, 27].
- **Between-individual biological variability**, which refers to the true differences across the homeostatic setting points of the individuals [26, 27]. Ideally, the variability between individuals should be high compared to any other aforementioned source of variability, as this indicates that the biomarker is reliable. ***Reliability*** is defined in this thesis as the ability of a biomarker to distinguish patients with a better test result (i.e., a better health outcome) from those with a worse test result (i.e., a worse health outcome), despite the presence of any measurement error [13, 15]. The higher the between-individual variability in relation to any other sources of variability (i.e., the measurement error), the higher the reliability of the biomarker.

Studies of physiologic and imaging tests

Studies of physiologic and imaging tests also aim to examine the true biological variability between individuals, as well as the inherent variability of the produced measurements [15]. Like laboratory-based tests, both physiologic and imaging tests can incur measurement error at the pre-analytical, analytical, and within-patient biological level. However, an additional source of measurement error, which is often of interest, includes that due to inter or intra-observer variability. ***Inter-observer variability*** refers to the systematic differences observed between two or more observers, when assessing the same observation (e.g., radiologists rating the same X-ray), while ***intra-observer***

variability refers to the systematic differences observed within the same observer, when assessing the same observation multiple times (e.g., the same radiologist giving a different rating when assessing the same X-ray twice) [15]. This source of error is usually a concern for imaging tests, where clinicians are required to read the produced images. In contrast, measurements of physiologic tests are produced directly from objective devices, with no clinical interpretation required.

1.3.4. Issues in the field of reliability and measurement error of biomarkers

Performing a meta-analysis of parameter estimates of reliability and measurement error

Systematic reviews are often conducted by searching the medical literature to identify primary studies, so that summary evidence is provided for research purposes, guideline development, evidence-based patient care and policy-making [15, 28-30]. Meta-analyses may in turn be performed within systematic reviews, and involve using statistical methods to synthesize quantitative evidence from primary studies, so that an overall estimate is obtained based on a whole body of research [31]. Although established in other areas of medical research (e.g., prognostic or diagnostic research, effectiveness of new interventions), methods for performing a meta-analysis of the measurement error of biomarkers are not well-developed.

Estimating the reliability and measurement error of count-based biomarkers

Statistical methods for estimating potential sources of measurement error in primary studies have been proposed, and are described in detail in Chapter 2. These methods assume underlying continuum and normality of the produced measurements at each potential level of variability. However, the assumption of normality is often violated when biomarkers are expressed as counts (i.e., whole numbers) rather than values on a continuous scale. Yet, no methodology has been proposed for examining the reliability and measurement error of count-based biomarkers.

1.4. Aim and objectives of the thesis

The broad aim of this thesis is to explore statistical issues around the evaluation of the reliability and measurement error of biomarkers, before biomarkers are used in medical research and practice. The overarching objectives of the thesis were:

i) to propose statistical methods for the meta-analysis of parameter estimates expressing the reliability and measurement error of continuous biomarkers, reported across primary studies.

ii) to propose statistical methods for the analysis of primary studies examining the reliability and measurement error of biomarkers expressed as counts rather than continuous measurements.

1.5. Thesis outline

This thesis broadly covers two areas, the meta-analysis of parameters of reliability and measurement error of continuous biomarkers, and the appropriate estimation of the reliability and measurement error of count-based biomarkers. In this thesis I have analysed case studies, reviewed and critiqued existing methods and developed new methods for the meta-analysis of parameters expressing the reliability and measurement error of continuous biomarkers, proposed and evaluated alternative methods for estimating the reliability and measurement error of count-based biomarkers in primary studies, and carried out simulation studies to examine how the reliability and measurement error changes across different simulated scenarios.

Description of standard statistical methods for primary analysis and meta-analysis, with application to a primary study

Chapter 2 provides the background on the state-of-the-art in the design and statistical analysis of primary studies examining the reliability and measurement error of continuous biomarkers, and the background on general methods used for the meta-analysis of estimates reported in primary studies.

Chapter 3 provides a detailed statistical analysis of a study examining the reliability and error of measurements of grip strength, produced from a digital dynamometer. For this purpose, data from patients with sarcopenia and chronic inflammatory disease were used as a case study. The aim of the chapter is to illustrate how standard methods proposed for estimating the reliability and measurement error of continuous biomarkers work, and to provide evidence of the reliability and measurement error of the Takei digital dynamometer, when used to evaluate grip strength.

Methods for the meta-analysis of parameters of reliability and measurement error

Chapter 4 is a methodological review of published systematic reviews reporting the reliability and measurement error of biomarkers evaluating the presence or progress of any pathological condition. The aim of the chapter is to appraise the review process used in the identified systematic reviews, and examine the current state of statistical methods used for the meta-analysis of parameters of reliability and measurement error.

In Chapter 5, the limitations of the meta-analytic methods used for continuous biomarkers, identified in Chapter 4, are discussed, and new methods for the meta-analysis of estimates of the reliability and measurement error of continuous biomarkers are proposed.

The methods proposed in Chapter 5 were in turn used in Chapter 6. A systematic review was carried out to identify primary studies examining the reliability and error of the grip strength measurements, produced for different from handheld dynamometers. A meta-analysis of the estimates of reliability and measurement error was then performed, using the methods presented in Chapter 5. The aim of the chapter is to illustrate how these meta-analytic methods are applied, and

to provide summary evidence of the error in the grip strength measurements based on a whole body of research.

Methods for estimating the reliability and measurement error of count-based biomarkers

Chapter 7 introduces alternative statistical methods for estimating reliability and measurement error when the measured response is expressed as a count, rather than in a continuous scale. The aim of the chapter is to present alternative methods that can be used for a different type of data, where the assumption of the underlying normality of the produced measurements is not valid.

In Chapter 8, the performance of the methods presented in Chapter 7 were compared to the standard methods used for estimating the reliability and error of continuous measurements. For this purpose, data from patients with Sjogren's syndrome who underwent labial salivary gland biopsy was used as a case study. The biomarker of interest was the focus score, calculated for each salivary gland observed in each biopsy.

Chapter 9 uses simulation to investigate the impact of different numbers of biopsy glands on the reliability and measurement error of the focus score.

Summary of findings

Chapter 10 summarises the findings and concludes the thesis. The chapter also discusses the strengths and limitations of the thesis, and recommends further work.

2. Background on primary studies examining the measurement error of biomarkers, and the meta-analysis of estimates reported in primary studies

2.1. Introduction

Before biomarkers are used in clinical practice, it is essential that researchers and medical professionals are familiar with any sources of error that may affect the reproducibility of the produced measurements. Primary studies have been proposed for this purpose. Systematic reviews may in turn collect any primary studies published in the medical literature. Meta-analysis is in turn often used to combine the estimates reported within primary studies, so that summary quantitative evidence of the measurement error of biomarkers is produced. This chapter provides i) the background on the state-of-the-art in the design, statistical analysis, and reporting of primary studies examining the measurement error of biomarkers; and ii) the background on general methods used for the meta-analysis of estimates reported in primary studies.

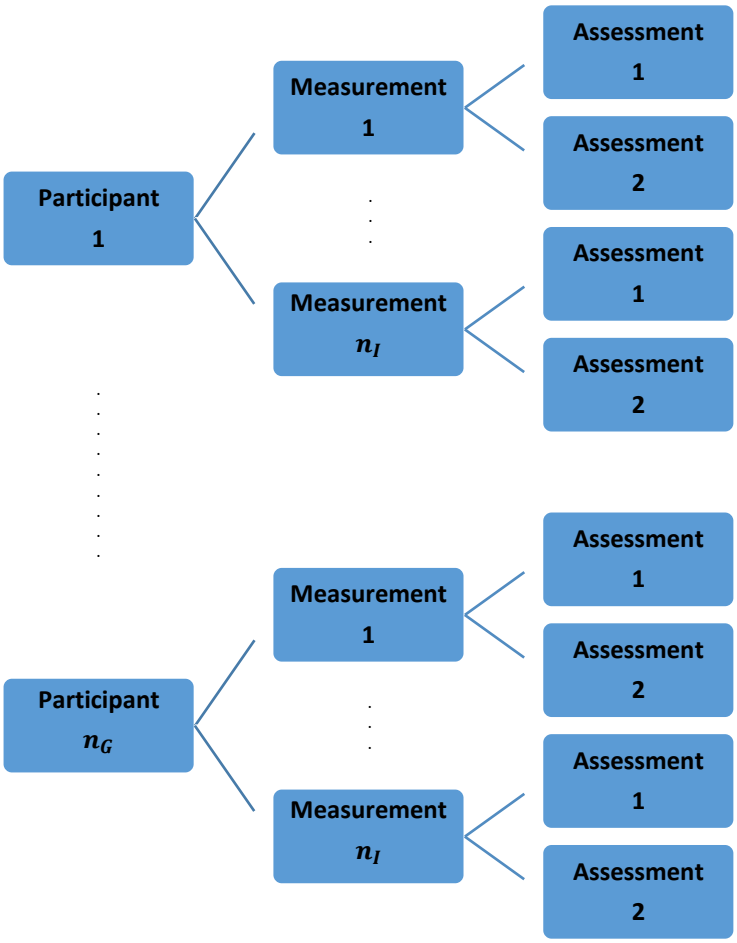
2.2. Design of primary studies examining the measurement error of biomarkers

This section discusses the issues around the design of studies examining the measurement error of biomarkers. When designing such studies, the issues considered are similar for all three types of tests (laboratory, imaging, physiologic), with minor alterations depending on the features of the test being examined.

The Fraser-Harris framework

Fraser and Harris provided a framework for the design of primary studies examining potential sources of error in the measurements of laboratory tests [26] (depicted in Figure 2.1). For laboratory tests, potential sources of measurement error include the within-individual biological and analytical variability in the measurements (introduced in Chapter 1). To assess the within-individual biological variability, multiple samples are taken from each individual at different time points, with the measurements produced from the analysis of the samples being compared to each other. The random analytical variability is in turn assessed by analysing the samples collected from each participant at the same time, in duplicate. This approach for assessing random analytical variability was introduced by Young et al [32], and has been advocated by researchers working extensively in the field of biological variability [17, 26, 27]. The main advantage of this approach is that any potential variability attributed to samples being analysed at different runs is eliminated. Furthermore, Fraser and Harris recommend that the analysis of the collected samples should be performed with a single instrument, a single set of calibrators, a single lot of reagents, and a single operator. This is to eliminate any potential systematic analytical variability arising between the duplicated measurements [26].

Figure 2.1. Design of biological variability study with a total of n_G participants and n_I measurements produced from each participant.



Tests where duplicate measurements are not possible

A similar study design has been adopted for primary studies examining the measurement error of physiologic or imaging tests, with the recruited individuals being assessed at multiple testing occasions. The measurements at two or more testing occasions may be taken either from the same clinician [33] or from different clinicians [34]. The measurement error is expected from any random variability in the measurements made within the individuals, at both the analytical and within-individual biological level. However, unlike laboratory-based tests, a formal assessment of the analytical variability is often not possible for physiologic tests (e.g., spirometry or blood pressure) and imaging tests (e.g., MRI or ultrasound), as the produced outcome cannot be assessed in duplicate. This means that, unlike laboratory tests, the amount of random variability attributed solely to the analytical imprecision of a physiologic or an imaging test cannot be quantified.

Tests with additional systematic error introduced by subjective interpretation

For some tests, an additional source of measurement error includes systematic variability between different observers (inter-observer) or within the same observer (intra-observer), when assessing the same observation multiple times. This source of error is more often of interest with imaging tests (e.g., two clinicians may rate the same image differently). In order to examine potential inter/intra-observer variability directly, the produced measurements are often assessed independently by multiple observers or twice by the same observer [35]. In the case of multiple observers assessing each measurement, observers may be deliberately chosen to represent both experienced and inexperienced observers [36].

2.2.1. Minimising pre-analytical variability prior to testing

Pre-analytical variability (introduced in Chapter 1) occurs due to differences across two or more testing occasions, in how an individual has prepared prior to each testing occasion, or how the within-individual measurements have been obtained or handled. Measurements should be taken to retain pre-analytical variability at a minimum, by keeping the testing conditions consistent. Thus, preparatory actions are required prior to performing the measurements, so that each measurement is standardised as much as possible. Any such actions should be clearly specified in a strictly implemented testing protocol [16, 17, 26, 37]. These may include the preparation of the patients (adhere to instructions on diet, rest, clothing or medication, undertake a familiarisation session if learning effects are likely to be present), the testing environment (light conditions, appropriate temperature), the professionals assessing the patients (undertake appropriate training, provide clear instructions to the individuals), and the device/equipment being used (calibration, adjustment of settings) [17, 26, 37].

Standardising the testing conditions is generally one of the most challenging parts of taking any measurement, is often underestimated, and may significantly affect the usefulness of medical interventions or tests [38, 39]. This source of variability may be of increased concern particularly with self-testing (which is becoming more and more common, particularly after the COVID-19 outbreak [40]), as the required pre-testing preparatory actions are likely to be performed in a less strict way compared to e.g., the laboratory setting. However, this source of variability is not a focus of the thesis, and is assumed to be minimised.

2.2.2. Populations investigated

There has been a long-standing debate on which patients should be recruited into laboratory-based studies. Fraser and Harris originally recommended that only apparently healthy

individuals should be considered since the main interest is the biological sources of variability, rather than pathological [17, 26]. More recently however, Fraser stated that valid estimates for the different components of variability can also obtained from diseased individuals, given that the status of the disease remains consistent [41, 42]. Braga and Panteghini recommend against the recruitment of diseased individuals, as it is difficult to define disease stability a priori [27].

A different view is expressed by De Vet et al [15]. The authors state that the selected individuals should reflect the population that is of interest. If for example it is of interest to know the potential error of measurements obtained from patients with a particular disease, then there is no use testing healthy individuals.

2.2.3. Time interval between measurements

For laboratory-based studies, Fraser and Harris state that the sample collection should be performed over a reasonably short period of time, so that the underlying disease status of the individual remains consistent, and valid estimates for the different components of variability are obtained [17, 26]. Braga and Panteghini recommend that samples should be collected at regular and fixed time intervals, with the authors additionally stating that the 'sample time interval and the study duration should be related to re-testing times used for the measurements of the specific analyte in clinical practice' [27].

De Vet et al state that there are no standard rules in choosing an appropriate time interval between two tests [15]. For diseases known to progress rapidly, the time interval between tests should be short, while diseases that are known to be stable may allow longer intervals to be used. The authors additionally state that there should be a good balance between the disease remaining stable, and the absence of any potential interferences (e.g., fatigue from the first measurement when physical activity is required from the individual).

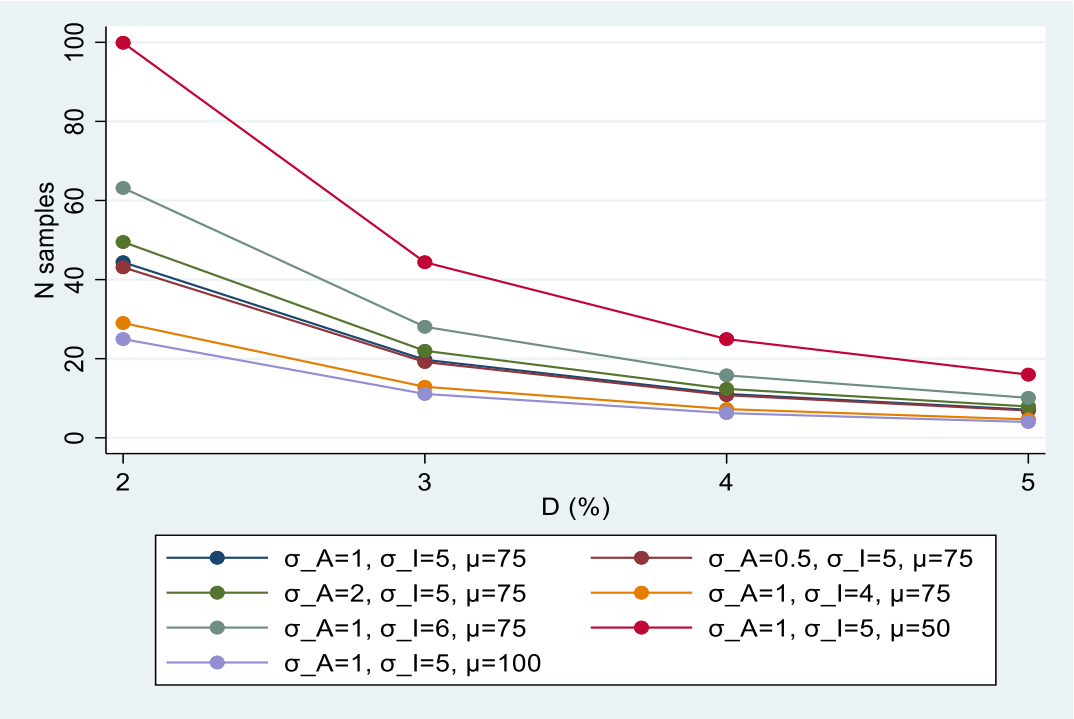
2.2.4. Sample size

There is limited guidance regarding the number of individuals and number of measurements from each individual required for designing a test variability study. Fraser and Harris [17, 26] state that there is no definite answer on how many individuals are required to conduct a laboratory-based variability study, as this decision is a compromise between a large number that would reduce uncertainty around the variability estimates, and a smaller number that will enable the samples to be handled appropriately and analysed under the right conditions. With respect to the number of samples, Braga and Panteghini [27] state that the higher the number collected from each individual, the more precise the estimate of the true value of the individual will be. The authors additionally provide a formula for evaluating the number of samples required from each individual, so that the produced mean value of multiple measurements lies within $D\%$ of the true value of the individual. The number of samples is derived from

$$N_{samples} = \frac{1.96^2 \left(\frac{\sigma_I^2 + \sigma_A^2}{\mu^2}\right)}{D^2}, \tag{2.1}$$

where σ_I and σ_A are the standard deviations of measurements produced at the within-individual biological and analytical levels, and μ is the grand mean of the measurements (parameters introduced in section 2.3.2.1). Figure 2.2 depicts how different pre-specified values of the aforementioned parameters affect the number of samples required for a produced mean of multiple within-individual measurement lying within $D\%$ of the true value of the individual, with a lower grand mean and larger variability estimates leading to a higher number of samples.

Figure 2.2. Impact of the grand mean, the analytical and within-individual biological standard deviations on the number of samples required for a mean of multiple within-individual measurements lying within $D\%$ of the true value of the individual.



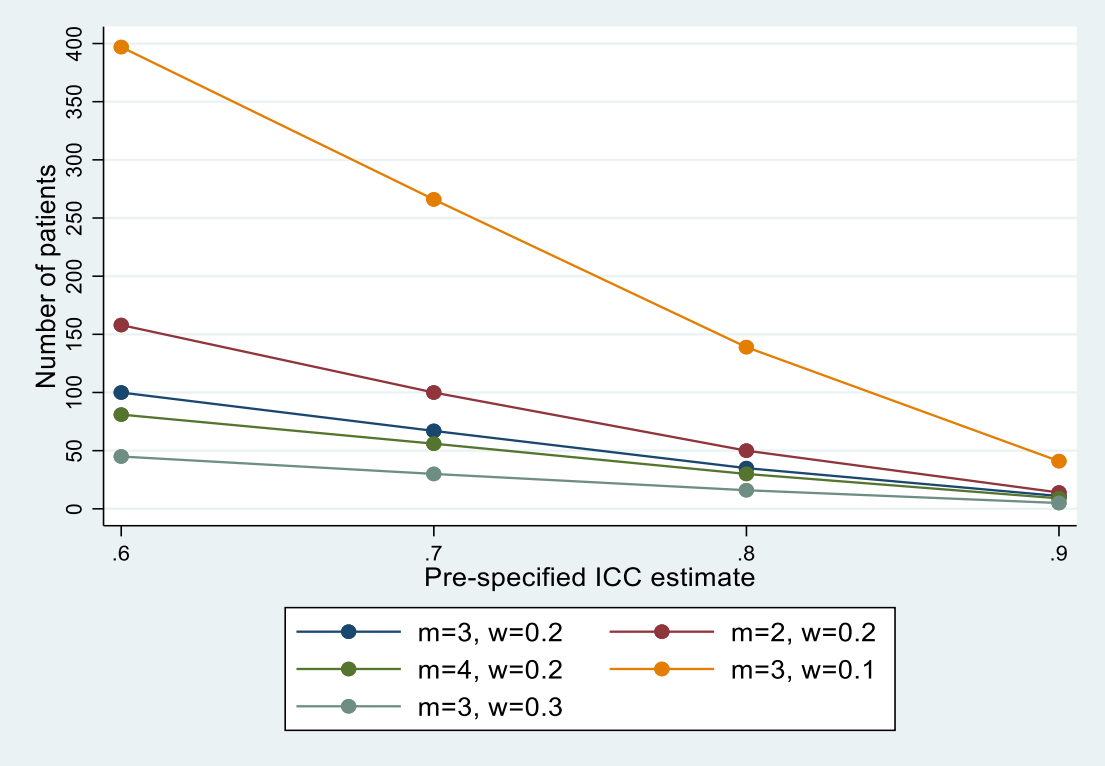
De Vet et al [15] suggest that 50 individuals are a sufficient for performing a variability study, and usually feasible to obtain. Giraudeau and Mary [43] provide a formula for estimating the number of individuals required, based on a number of measurements per individual, a pre-specified estimate for the intra class correlation (parameter introduced in section 2.3.3.3), and a pre-specified width for the 95% confidence interval of the estimated intra class correlation. Using this formula, the sample size is calculated as

$$N_{patients} = \frac{8 \times 1.96^2 (1 - ICC)^2 [1 + (m-1)ICC]^2}{m(m-1)w^2}, \tag{2.2}$$

where m is the number of measurements per individual, ICC is a pre-specified estimate for the intra class correlation, and w is a pre-specified width for the 95% confidence interval of the ICC estimate. Figure 2.3 shows how different pre-specified values of m, ICC, and w impact the

required number of individuals. As depicted in Figure 2.3, a higher value of m, ICC, and w appears to significantly reduce the required number of individuals.

Figure 2.3. Impact of the number of within-individual measurements and the pre-specified ICC and 95% confidence interval on the number of individuals required.



2.3. Statistical analysis of primary studies examining the reliability and measurement error of biomarkers

This section presents the current state of statistical methods used in primary studies examining potential sources of variability in the measurements of biomarkers. Except for the Kappa statistic, which is used in situations where the produced measurements are categorical (binary or ordinal), all methods described in this chapter assume underlying continuum and normality of the measurements, at all potential levels of variability. No statistical methods

were identified for measurements expressed as counts (i.e., whole numbers) rather than continuous.

If available, methods for constructing a 95% confidence interval for each parameter are also presented. Furthermore, if applicable, the chapter provides methods for estimating each parameter when the measurements are log-transformed prior to analysis, which is a common approach used when the normality assumption of the measurements is violated.

2.3.1. Preparation of the data prior to analysis

Prior to performing a statistical analysis, studies are often concerned with several data assumptions that are required for the use of standard statistical methods for estimating the potential sources of test variability. Such assumptions may include the absence of any significant outliers, the normality of the measurements at each potential level of variability, and the homoscedasticity across different individuals.

Outliers

In studies of laboratory tests, it is recommended that outliers are removed prior to the analysis, as even a single outlier may remarkably influence the estimation of the different components of variability [17, 27]. Braga and Panteghini provide clear guidance on how the detection of outliers should be carried out (see Figure 2.4) [27].

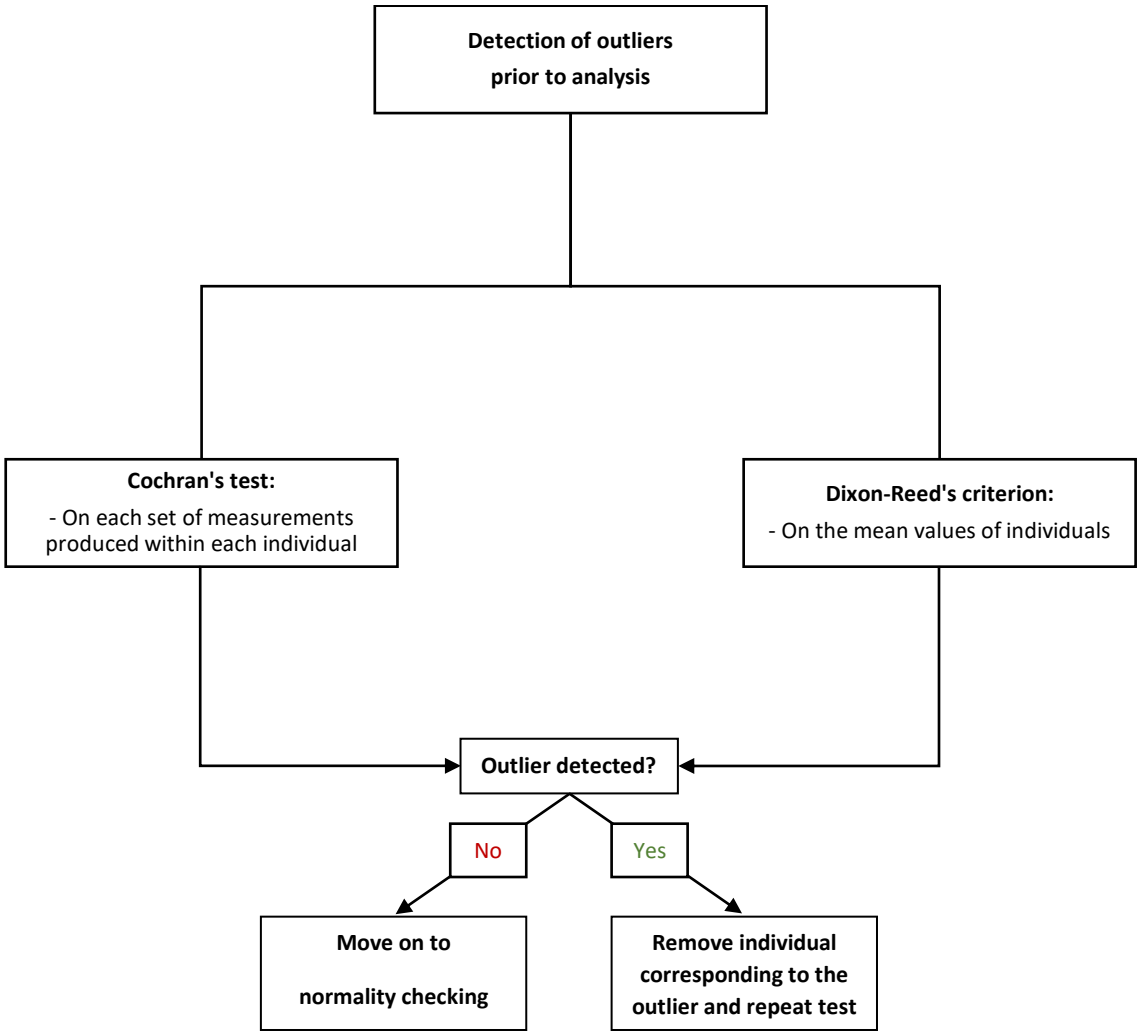
The authors suggest the use of the Cochran's test [44] to examine the presence of any outliers (i) in the measurements produced within individuals, and (ii) in the duplicate results of each measurement. Assuming that all individuals produce the same number of measurements, and that the measurements are normally distributed around the true value of the individual, this test examines whether a variance estimate from a single individual is significantly larger

compared to all other variance estimates produced from the remaining individuals. If significantly larger, it is recommended that all measurements produced from this individual should be excluded.

The authors then recommend using the Dixon-Reed's criterion [45] to assess whether the mean value of the measurements produced from each individual is an outlier. This test considers the difference between an extreme value and the next lowest (or highest) value, and rejects the extreme value if the difference exceeds one-third of the range of all values.

An alternative view on handling outliers is expressed by De Vet et al [15]. The authors state that outliers should not be removed, as they do occur in real life, and may indicate difficulties when performing a measurement (e.g., clinicians reading out a measurement from an image) [15].

Figure 2.4. Detection of outliers as described in Braga and Panteghini [27].



Normality of the data

For studies of laboratory tests, Fraser and Harris [17, 26] recommend the use of Shapiro-Wilk test [46] to examine both the distribution of the measurements produced within the individuals, and the distribution of the mean values obtained from each individual. More detailed guidance on how normality should be checked was subsequently provided by Braga and Panteghini (depicted in Figure 2.5) [27]. However, a formal assessment of normality at each different level may often not be reasonable (or even possible) due to limited numbers of observations (e.g., only two measurements taken from each individual).

Normality is also desired when measurements are produced from physiologic or imaging tests (given that the outcome produced from reading an image is a score rather than e.g., a binary or ordinal response), with the log-transformation being frequently applied in case the measurements are skewed [16]. However, unlike studies of laboratory tests, no specific guidelines have been provided on how normality should be checked.

Figure 2.5. Normality checking as described in Braga and Panteghini [27].

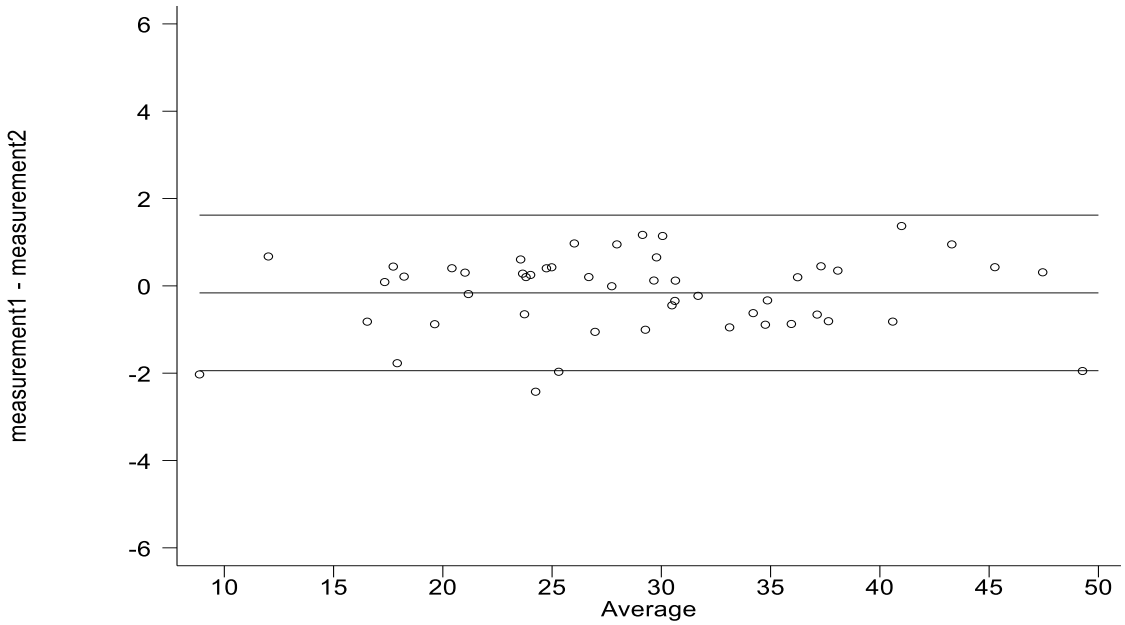
[1] "Yes" *means that the normality assumption is met, while* "No" *means that the normality assumption is violated.*

Homoscedasticity of the data

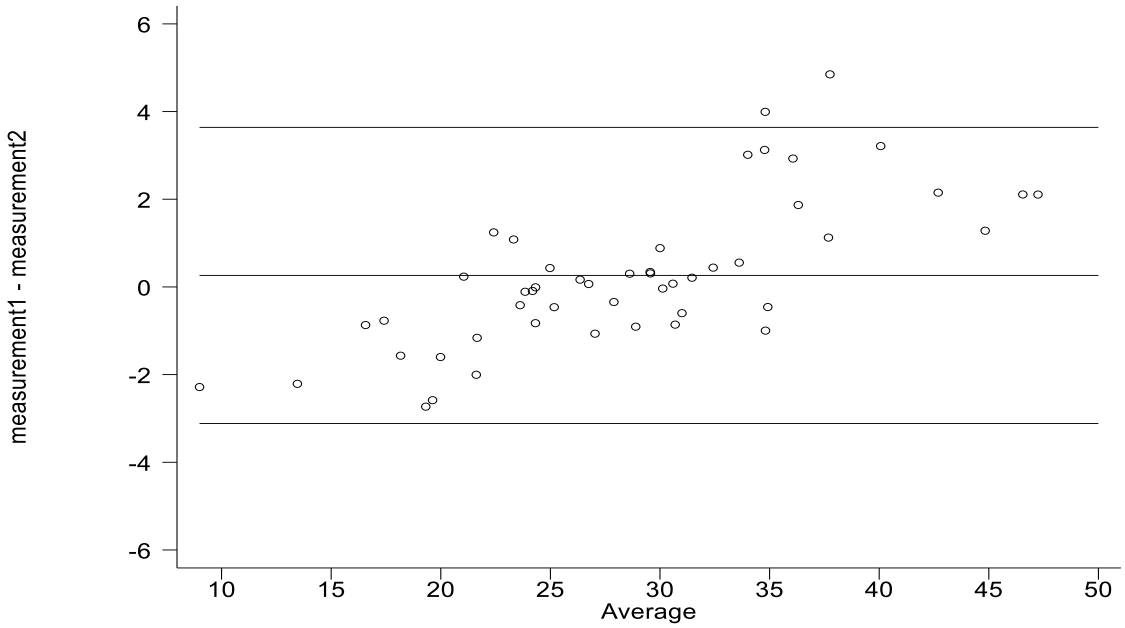
In studies of non-laboratory tests, an additional assumption which is often of concern is that of homoscedasticity [16]. That is, the variability in the measurements produced within each individual should be the same across all individuals recruited in the study [16, 47]. When two measurements are taken from each individual, the presence of heteroscedasticity in the data is recommended to be examined through a Bland-Altman plot [16]. Bland-Altman plots are constructed by plotting the absolute difference between two measurements produced within each individual, against the corresponding mean of the measurements [47]. If heteroscedasticity appears present (see Figure 2.6 for an example), it is recommended that data are log-transformed prior to analysis. In studies of laboratory tests, there is no guidance for checking this assumption [16, 47].

Figure 2.6. Two examples of Bland-Altman plots of the difference between two measurements produced from the same individual against the corresponding average value.

i) *Homoscedasticity assumption met.* The within-individual variability does not change with higher average values of the test.



ii) *Homoscedasticity assumption violated.* The within-individual variability increases for higher average values of the test.



2.3.2. General method used for estimating sources of variability in the measurements

For all types of tests, it is recommended that the multiple sources of variability in the measurements are estimated through a nested analysis of variance (ANOVA) model. Given that the number of observations within individuals is the same for all individuals, this method is equivalent to fitting a linear regression random effects model. These models assume that the measurements are expressed in a continuous scale. The models may take a different form, depending on the type of test being examined.

2.3.2.1. Studies of laboratory tests

The two-way nested ANOVA is the established method used for the analysis of biological variability studies. The use of this method was originally proposed by Fraser and Harris [26]. Using this method, the variance at the analytical, within-individual, and between-individual biological level is estimated as shown in Table 2.1. When the number of observations (measurements and assessments per

measurement) are the same for each individual, this method is equivalent to fitting a linear regression model with random effects only. The model is expressed as

$$y_{ijk} = \mu + G_k + I_{jk} + A_{ijk}, \quad (2.3)$$

where y_{ijk} denotes the i_{th} assessment ($i = 1,2$) of the j_{th} measurement ($j = 1, \dots, n_I$) produced from the k_{th} individual ($k = 1, \dots, n_G$) , and μ is the regression intercept. $G_k \sim N(0,\ \sigma_G)$ is the group-level random effects parameter for the true variability across the individuals, $I_{jk} \sim N(0,\ \sigma_I)$ is the individual-level random effects parameter for the within-individual biological variability, and $A_{ijk} \sim N(0,\ \sigma_A)$ is the random error term for the analytical variability, with all three parameters normally distributed with zero mean and standard deviation of σ_G, σ_I, and σ_A, respectively. The four parameter estimates obtained from the above regression model are:

- The grand mean of the produced measurements, which equals the regression intercept (μ).
- The estimate of the standard deviation for the between-individual biological variability (σ_G).
- The estimate of the standard deviation for the within-individual biological variability (σ_I).
- The estimate of the standard deviation for the analytical variability (σ_A).

When using a random effects linear regression model to estimate each component of variability, the use of restricted maximum likelihood (REML) is recommended, as this method yields less biased estimates compared to the standard maximum likelihood approach, particularly with small sample sizes [48].

Table 2.1. ANOVA-based calculations for estimating the components of variability in laboratory tests.

Variance component	Degrees of freedom	Sum of squares	Mean square	Variance estimates
Between individuals	$n_G - 1$	$SS_G = n_A n_I \sum_{k=1}^{n_G} (\overline{y_k} - \bar{y})^2$	$MS_G = \frac{SS_G}{n_G - 1}$	$\sigma_G^2 = \frac{MS_G - MS_I}{n_I n_A}$
Within individuals	$(n_I - 1)n_G$	$SS_I = n_A \sum_{j=1}^{n_I} \sum_{k=1}^{n_G} (\overline{y_{Jk}} - \overline{y_k})^2$	$MS_I = \frac{SS_I}{(n_I - 1)n_G}$	$\sigma_I^2 = \frac{MS_I - MS_A}{n_A}$

Within assessments	$(n_A - 1)n_G n_I$	$SS_A = \sum_{i=1}^{n_A}\sum_{j=1}^{n_I}\sum_{k=1}^{n_G}(y_{ijk} - \overline{y_{jk}})^2$	$MS_A = \frac{SS_A}{(n_A - 1)n_G n_I}$	$\sigma_A^2 = MS_A$
TOTAL	$n_G n_I n_A - 1$	$\sum_{i=1}^{n_A}\sum_{j=1}^{n_I}\sum_{k=1}^{n_G}(y_{ijk} - \bar{y})^2$		$\sigma_G^2 + \sigma_I^2 + \sigma_A^2$

[1] $\bar{y} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_I}\sum_{k=1}^{n_G} y_{ijk}}{n_A n_I n_G}$, [2] $\overline{y_{jk}} = \frac{\sum_{i=1}^{n_I} y_{ijk}}{n_I}$, [3] $\overline{y_k} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_I} y_{ijk}}{n_A n_I}$

2.3.2.2. Studies of non-laboratory tests

Shrout and Fleiss provide guidance on the analysis of studies examining the variability of tests [49]. The authors present three different models for estimating the different sources of variability in the measurements, which should be used under different circumstances. Information on how each model is expressed and how each variability component is estimated, is presented in Table 2.2.

One-way random effects model

This model should be employed when potential differences between the measurements produced within individuals arise solely from random variability at the analytical and within-patient biological level. This is mostly encountered with physiologic tests, where measurements are produced from objective devices, and thus the effect of clinicians is expected to be negligible. The model is expressed as

$$y_{ij} = \mu + G_j + e_{ij}, \qquad (2.4)$$

where y_{ijk} denotes the i_{th} measurement $(i = 1, \dots, n_I)$ produced from the j_{th} individual $(j = 1, \dots, n_G)$, and μ is the regression intercept. $G_j \sim N(0,\ \sigma_G)$ is the group-level random effects parameter for the true variability across the individuals, and $e_{ij} \sim N(0,\ \sigma_{I+A})$ is a random error term for the remainder variability observed within the individuals. Both parameters are normally

distributed with zero mean and standard deviation of σ_G and σ_{I+A}, respectively. The three parameter estimates obtained from the above regression model are:

- The grand mean of the produced measurements, which equals the regression intercept (μ).
- The estimate of the standard deviation for the between-individual biological variability (σ_G).
- The estimate of the standard deviation corresponding to the remainder variability observed within the individuals (σ_{I+A}). This estimate combines within-individual biological and analytical variability, as in contrast to laboratory-based tests, a direct assessment of the analytical variability is not possible for non-laboratory tests.

Two-way random effects model

When systematic variability between different clinicians (inter-observer) or within the same clinician (intra-observer) performing the measurements is likely to occur and is of interest (which is nearly always the case), a two-way model should be the model of choice. This is mostly encountered with imaging tests, where a clinical interpretation of the produced images is often required for obtaining the measurements. This model assumes that the selected clinician(s) is a random sample of all possible clinicians possessing similar characteristics. Compared to the one-way model, this model is extended to

$$y_{ij} = \mu + G_j + O_i + e_{ij}, \qquad (2.5)$$

where μ, G_j, and e_{ij} are the same as in equation 2.4, and the additional random effects parameter $O_i \sim N(0,\ \sigma_O)$ estimates the standard deviation corresponding to potential inter or intra-observer variability. The four parameter estimates obtained from the above regression model are:

- The grand mean of the produced measurements, which equals the regression intercept (μ).
- The estimate of the standard deviation for the between-individual biological variability (σ_G).
- The estimate of the standard deviation for the between or within-observer variability (σ_O).

- The estimate of the standard deviation corresponding to the remainder variability observed within the individuals (σ_{I+A}), which again is a composite of the within-individual biological and analytical variability.

Two-way mixed effects model

This model should only be used only when the consistency (i.e., agreement in ranking individuals) is of interest, rather than the absolute agreement between different observes (or within the same observer). An example includes assigning priority scores to patients on a waiting list for heart surgery. In this case, one would only be interested in whether two clinicians rank the patients in the same order, based on severity of disease. In contrast to the random effects model, the results produced from a mixed effects model only represent the clinician(s) selected in the study, and cannot be generalised across all clinicians with similar characteristics. Under a two-way mixed effects model, the parameter O_i in equation 2.5 reduces to a fixed effect parameter. However, this model is rarely used for the analysis of test variability studies, as potential measurement error due to the absolute disagreement between or within clinicians is nearly always of interest.

Table 2.2. Methods for estimating the components of variability in non-laboratory tests.

Model	Notation	Variance component	Degrees of freedom	Sum of squares	Variance estimate
One-way random effects	$y_{ij} = \mu + G_j + e_{ij}$, where $G_j \sim N(0, \sigma_G)$ $e_{ij} \sim N(0, \sigma_{I+A})$	Between individuals	$n_G - 1$	$SS_G = n_I \sum_{j=1}^{n_G} (\bar{y}_J - \bar{y})^2$	$\sigma_G^2 = \frac{MS_G - MS_I}{n_I}$
		Within individuals	$(n_I - 1)n_G$	$SS_e = \sum_{i=1}^{n_I} \sum_{j=1}^{n_G} (y_{ij} - \bar{y}_J)^2$	$\sigma_{I+A}^2 = MS_I$
Two-way random effects	$y_{ij} = \mu + G_j + O_i + e_{ij}$, where $G_j \sim N(0, \sigma_G)$ $e_{ij} \sim N(0, \sigma_{I+A})$ $O_i \sim N(0, \sigma_O)$	Between individuals	$n_G - 1$	$SS_G = n_I \sum_{j=1}^{n_G} (\bar{y}_J - \bar{y})^2$	$\sigma_G^2 = \frac{MS_G - MS_e}{n_I}$
		Within individuals	$(n_I - 1)n_G$	$SS_I = \sum_{i=1}^{n_I} \sum_{j=1}^{n_G} (y_{ij} - \bar{y}_J)^2$	$\sigma_O^2 + \sigma_e^2 = MS_I$
		- *Between/within observers*	$n_I - 1$	$SS_O = n_G \sum_{i=1}^{n_I} (\bar{y}_i - \bar{y})^2$	$\sigma_O^2 = \frac{MS_O - MS_e}{n_G}$
		- *Random error*	$(n_I - 1)(n_G - 1)$	$SS_e = \sum_{i=1}^{n_I} \sum_{j=1}^{n_G} (y_{ij} - \bar{y}_J - \bar{y}_i + \bar{y})^2$	$\sigma_e^2 = MS_e$
Two-way mixed effects	$y_{ij} = \mu + G_j + O_i + e_{ij}$, where $G_j \sim N(0, \sigma_G)$ $e_{ij} \sim N(0, \sigma_{I+A})$ $\sum_{i=1}^{n_I} O_i = 0$	Between individuals	$n_G - 1$	$SS_G = n_I \sum_{j=1}^{n_G} (\bar{y}_J - \bar{y})^2$	$\sigma_G^2 = \frac{MS_G - MS_e}{n_I}$
		Within individuals	$(n_I - 1)n_G$	$SS_I = \sum_{i=1}^{n_I} \sum_{j=1}^{n_G} (y_{ij} - \bar{y}_J)^2$	$\sigma_O^2 + \sigma_e^2 = MS_I$
		- *Between/within observers*	$n_I - 1$	$SS_O = n_G \sum_{i=1}^{n_I} (\bar{y}_i - \bar{y})^2$	$\sigma_O^2 = \frac{\sum_{i=1}^{n_I} O_i^2}{n_I - 1}$
		- *Random error*	$(n_I - 1)(n_G - 1)$	$SS_e = \sum_{i=1}^{n_I} \sum_{j=1}^{n_G} (y_{ij} - \bar{y}_J - \bar{y}_i + \bar{y})^2$	$\sigma_e^2 = MS_e$

[1] MS is mean squares, estimated for each variance component as **(Sum of squares)** divided by **(Degrees of freedom),**

[2] $\overline{y_i} = \frac{\sum_{j=1}^{n_G} y_{ij}}{n_G}$, [3] $\overline{y_J} = \frac{\sum_{i=1}^{n_I} y_{ij}}{n_I}$, [4] $\overline{y} = \frac{\sum_{i=1}^{n_I} \sum_{j=1}^{n_G} y_{ij}}{n_I n_G}$

2.3.3. Statistical parameters reported in primary studies examining measurement error

Several parameters of measurement error can in turn be derived from the estimates produced from the aforementioned models. Such parameters include the standard error of measurement, the smallest detectable change, the intra-class correlation, the coefficient of variation, the index of individuality, and the reference change values. Furthermore, additional parameters commonly reported in studies of non-laboratory tests include the limits of agreement, the Pearson correlation coefficient, and the Kappa coefficient. Details of all parameters reported in this section is provided in Table 2.3. If available, methods for constructing a 95% confidence interval were also presented for each parameter. Furthermore, this section provides alternative methods for calculating each parameter when the measurements require log-transformation.

Table 2.3. Commonly reported parameters of reliability and measurement error.

Parameter	Formula
Standard error of measurement using a two-way random effects model[1]	$SEM = \sqrt{\sigma_i^2 + \sigma_j^2}$, where i, j denote two different levels of measurement error.
Standard error of measurement using a one-way random effects or a two-way fixed effect model[2]	$SEM = \sigma_{I+A}$, where σ_{I+A} is the estimated standard deviation for the combined analytical and within-individual biological variability.
Smallest detectable change[1]	$SDC = \sqrt{2} \times 1.96 \times SEM$
Intra class correlation[1]	$ICC = \frac{\sigma_G^2}{\sigma_G^2 + SEM^2}$, where σ_G is the estimated standard deviation for the between-individual biological variability.
Coefficient of variation[1]	$CV = \frac{SEM}{\mu}$, where μ is the grand mean of the measurements.

Index of individuality	$$II = \frac{\sqrt{\sigma_A^2 + \sigma_I^2}}{\sigma_G},$$ where $\sigma_A, \sigma_I, \sigma_G$ is the estimated standard deviation for the analytical, within and between-individual biological variability.
Reference change values	$$RCV = \sqrt{2} \times Z \times \frac{\sqrt{\sigma_A^2 + \sigma_I^2}}{\mu},$$ where Z is a *Z*-score selected from the normal distribution. *Z*=1.96 (=2.77) is used to assess whether a significant (highly significant) change in the health status of the patient has taken place.
Limits of agreement[1]	$$LoA = \bar{d} \pm 1.96 \times SD_{\bar{d}},$$ where $\bar{d}$ is the mean difference between two within-individual measurements, and $SD_{\bar{d}}$ is the corresponding standard deviation.
Pearson correlation coefficient	$$r = \frac{Cov(y_1, y_2)}{\sigma_{y_1}\sigma_{y_2}},$$ where y_1, y_2 are two measurements produced within the individuals at two different testing sessions, and $\sigma_{y_1}, \sigma_{y_2}$ are the standard deviations of each session.
Kappa coefficient[1]	$$K = \frac{p_o - p_e}{1 - p_e},$$ where p_o is the proportion of cases where inter/intra-observer agreement is achieved, and p_e the proportion of cases where the agreement is expected by chance.

[1] parameter appropriate to use when inter/intra-observer variability between measurements is present.

[2] parameter equals the SEM based on two-way random effects when inter/intra-observer variability between measurements is absent.

2.3.3.1. The standard error of measurement (SEM)

The standard error of measurement is equal to the standard deviation of multiple measurements produced within an individual, which reflects the spread of the produced measurements around the true score of the individual [14, 15]. The use of this parameter requires the following two assumptions to be met: (i) the measurements produced within individuals are sampled from a normal distribution, and (ii) there should be no evidence of heteroscedasticity in the data. Assuming normality of the measurements made within individuals and homoscedasticity across individuals, the standard error of measurement is the value up to which the absolute difference between a produced measurement and the true value of an individual is expected to lie with 68% probability [14]. This parameter is commonly reported in both laboratory and non-laboratory studies. For studies conducted in the laboratory setting, the standard error of measurement (termed as "the total error" by Fraser [17]) is calculated as

$$SEM = \sqrt{\sigma_I^2 + \sigma_A^2}\,, \tag{2.6}$$

where σ_I^2 and σ_A^2 are the variance estimates for the analytical and within-individual biological components, respectively. This reflects the total error that is expected due to variability arising at both the within-individual biological and analytical levels.

Similarly, for non-laboratory tests, the standard error of measurement is calculated as

$$SEM = \sqrt{\sigma_O^2 + \sigma_{I+A}^2}, \tag{2.7}$$

where σ_O^2 is the variance due to systematic differences between observers (or within the same observer), and σ_{I+A}^2 is the variance including both the analytical and within-individual biological variability. However, when systematic differences between or within-observers are either expected to be negligible, or not particularly of interest, the calculation reduces to

$$SEM = \sigma_{I+A} \tag{2.8}$$

Standard error of measurement for averaged measurements

It is known that measurements produced from specific tests (e.g., blood pressure) are substantially variable, either because of natural fluctuations, or because of the way the measurements are performed by the clinicians. When examining the reproducibility of such tests, it is common practice to take multiple within-individual measurements at each testing occasion and use the mean of these repeated measurements as a summary measure. This is based on the assumption that the average of multiple measurements performed within an individual is expected to be closer to the true value of the individual, compared to a single measurement [15]. Under this situation, the calculation of the standard error of measurement should adjust for the use of the mean as a summary measure. This is done by

$$SEM_k = \frac{SEM}{\sqrt{k}}, \tag{2.9}$$

where SEM is estimated as described above, and k is the number of measurements taken from each individual at each testing occasion [15]. Dividing by $\sqrt{k}$ accounts for the fact that the calculation of the average is based on k measurements. Each is accompanied by error, which would be incorporated k times in the estimate of SEM if $\sqrt{k}$ was not present in equation 2.9.

95% confidence intervals for the standard error of measurement

When the standard error of measurement is estimated through equation 2.8, a 95% confidence interval is obtained as

$$\left[\sqrt{\frac{SEM^2(n_I-1)n_G}{\chi^2_{(n_I-1)n_G,\ 0.975}}}, \sqrt{\frac{SEM^2(n_I-1)n_G}{\chi^2_{(n_I-1)n_G,\ 0.025}}}\right], \tag{2.10}$$

where n_G is the number of recruited individuals, n_I is the number of measurements taken from each individual, and $\chi^2_{(n_I-1)n_G,\ a}$ is the $(a^{th} \times 100)$ centile of the Chi-square distribution with $(n_I - 1)n_G$ degrees of freedom [13].

When the standard error of measurement is expressed as a linear function of two independent variance estimates (as in equations 2.6 and 2.7), a confidence interval can be obtained using the methods proposed by Graybill and Wang [50]. If $\gamma = \sigma_1^2 + \sigma_2^2$, and σ_1^2, σ_2^2 are two mutually independent variance estimates with

$$\frac{q_i \sigma_i^2}{E[\sigma_i^2]} \sim \chi^2_{q_i}, \tag{2.11}$$

where $i = 1,2$, q_i is a known integer, and $\chi^2_{q_i}$ denotes the chi-square distribution with q_i degrees of freedom, then a 95% confidence interval for $\sqrt{\gamma}$ can be obtained as

$$\left[\sqrt{\gamma - \sum_{i=1}^{2} G_i^2 \sigma_i^2}\ ,\ \sqrt{\gamma + \sum_{i=1}^{2} H_i^2 \sigma_i^2}\right], \tag{2.12}$$

where

$$G_i^2 = 1 - \frac{1}{F_{0.025:\ q_i,\infty}},\quad H_i^2 = \frac{1}{F_{0.975:\ q_i,\infty}} - 1 \tag{2.13}$$

and $F_{a:\ q,p}$ is the upper a percentage point of the F distribution with q and p degrees of freedom.

Finally, when averaged measurements are used as a summary measure, the lower and upper confidence bounds of SEM should be divided by $\sqrt{k}$ (as with the original estimate).

Calculation of the standard error of measurement for log-transformed data

When the measurements are log-transformed, the calculation of the standard error of measurement using the methods described above becomes meaningless. As this parameter is expressed on the original measurement scale, there is no more a natural interpretation of the produced estimate, and no methods have been proposed for reverting the produced estimate to the original scale.

2.3.3.2. The smallest detectable change (SDC)

The smallest detectable change can in turn be calculated as

$$SDC = \sqrt{2} \times 1.96 \times SEM, \tag{2.14}$$

where SEM is the standard error of measurement (calculated as described in section 2.3.3.1). The smallest detectable change reflects the quantity below which the absolute difference two measurements produced from the same individual is expected to lie with 95% probability due to measurement error, and not due to a true change in the performance of the individual [15, 51].

95% confidence intervals for the smallest detectable change

Given that the smallest detectable change is directly related the standard error of measurement, confidence intervals for this parameter can be obtained as

$$\left[\sqrt{2} \times 1.96 \times SEM_L, \sqrt{2} \times 1.96 \times SEM_U\right], \tag{2.15}$$

where SEM_L and SEM_U denote the upper and lower confidence bounds of the standard error of measurement, respectively.

Calculation of the smallest detectable change for log-normal data

Alike the standard error of measurement, no methods are available for calculating the smallest detectable change when measurements are log-transformed.

2.3.3.3. The intra class correlation (ICC)

The intra class correlation, also known as the reliability parameter, expresses how reliable the measurements produced from a specific test are. ***Reliability*** refers to the ability of a test to distinguish patients from each other despite the presence of measurement error [13, 15]. This parameter represents the proportion of the total variability in the measurements that is attributed to true differences between the patients. When potential variability between or within observers is expected to be negligible, or not particularly of interest, the parameter is calculated as

$$ICC = \frac{\sigma_G^2}{\sigma_{Total}^2} = \frac{\sigma_G^2}{\sigma_G^2 + SEM^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{I+A}^2}, \tag{2.16}$$

where σ_G^2 is the variance attributed to true differences between patients, and σ_{I+A}^2 is the variance including both the analytical and within-patient biological variability. If potential variability between or within observers is considered, the calculation is extended to

$$ICC = \frac{\sigma_G^2}{\sigma_{Total}^2} = \frac{\sigma_G^2}{\sigma_G^2 + SEM^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_O^2 + \sigma_{I+A}^2}, \tag{2.17}$$

where σ_O^2 is the variance attributed to true differences between observers (or within the same observer when rating the same individual multiple times). Guidelines for the interpretation of the intra class correlation have been provided by Cicchetti [52] and Koo and Li [53] (see Table 2.4), with values >0.90 accepted as indicative of excellent reliability.

In addition to evaluating the test reliability, this parameter is commonly used in cluster randomised trials to express the proportion of the total variability that is attributable to systematic differences between clusters, with the term ‘clusters’ usually referring to the recruiting medical centres [54].

Intra class correlation for averaged measurements

When the mean of multiple within-individual measurements performed at each testing occasion is used as a summary measure, the calculation is further extended to

$$ICC = \frac{\sigma_G^2}{\sigma_{Total}^2} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{SEM^2}{k}}, \tag{2.18}$$

where k denotes the number of measurements taken from each individual at each testing occasion [15].

Table 2.4. Interpretation of the intra class correlation based on Cicchetti [52], and Koo and Li [53].

Cicchetti [52]			Koo and Li [53]	
ICC less than 0.40	poor		ICC less than 0.50	poor
ICC between 0.40 and 0.59	fair		ICC between 0.50 and 0.74	moderate
ICC between 0.60 and 0.74	good		ICC between 0.75 and 0.90	good
ICC higher than 0.75	excellent		ICC higher than 0.90	excellent

95% confidence intervals for the intra class correlation

Shrout and Fleiss [49, 55] derived approximate confidence intervals for cases were the estimated intra class correlation additionally accounts for variance due to observers, while exact confidence intervals are available when the potential variance due to observers is not considered. The methods for calculating a lower and upper 95% confidence bound for each different form of the intra class correlation are presented in Table 2.5. The calculations are based on the F-distribution.

Table 2.5. Methods for calculating 95% confidence intervals for all different types of intra class correlations.

Formula	Lower 95% confidence bound	Upper 95% confidence bound
- *ICC from one-way random effects (single measurement)*		
$ICC = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{I+A}^2}$	$\frac{F_L - 1}{F_L + k - 1}$ where $F_L = F_{obs}/F_{0.975}(n-1, n\times(k-1)),\ \ F_{obs} = \frac{MS_G}{MS_E}$	$\frac{F_U - 1}{F_U + k - 1}$ where $F_L = F_{obs}\times F_{0.975}(n\times(k-1), n-1),\ \ F_{obs} = \frac{MS_G}{MS_E}$
- *ICC from one-way random effects (mean of multiple measurements)*		
$ICC_k = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{I+A}^2}{k}}$	$1 - \frac{1}{F_L}$ where F_L is defined as above	$1 - \frac{1}{F_U}$ where F_L is defined as above
- *ICC from two-way mixed effects (single measurement)*		
$ICC = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{I+A}^2}$	$\frac{F_L - 1}{F_L + k - 1}$ where $F_L = F_{obs}/F_{0.975}(n-1, (n-1)\times(k-1)),\ \ F_{obs} = \frac{MS_G}{MS_E}$	$\frac{F_U - 1}{F_U + k - 1}$ where $F_L = F_{obs}\times F_{0.975}(\ (n-1)\times(k-1), n-1),\ \ F_{obs} = \frac{MS_G}{MS_E}$
- *ICC from two-way mixed effects (mean of multiple measurements)*		
$ICC_k = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{I+A}^2}{k}}$	$1 - \frac{1}{F_L}$ where F_L is defined as above	$1 - \frac{1}{F_U}$ where F_L is defined as above
-		

Formula	Lower 95% confidence bound	Upper 95% confidence bound
- *ICC from two-way random effects (single measurement)*		
$ICC = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_O^2 + \sigma_{I+A}^2}$	$\frac{n \times [MS_G - F_{0.975}(n-1, v) \times MS_E]}{F_{0.975}(n-1, v) \times [k \times MS_O + (k \times n - k - n) \times MS_E] + n \times MS_G}$ where $v = \frac{(a \times MS_O + b \times MS_E)^2}{\frac{(a \times MS_O)^2}{k-1} + \frac{(b \times MS_E)^2}{(n-1) \times (k-1)}}$ and $a = \frac{k \times \widehat{ICC}}{n \times (1 - \widehat{ICC})}, \quad b = 1 + \frac{k \times \widehat{ICC} \times (n-1)}{n \times (1 - \widehat{ICC})}$	$\frac{n \times [F_{0.975}(v, n-1) \times MS_G - MS_E]}{k \times MS_O + (k \times n - k - n) \times MS_E + n \times F_{0.975}(v, n-1) \times MS_G}$ where $v = \frac{(a \times MS_O + b \times MS_E)^2}{\frac{(a \times MS_O)^2}{k-1} + \frac{(b \times MS_E)^2}{(n-1) \times (k-1)}}$ and $a = \frac{k \times \widehat{ICC}}{n \times (1 - \widehat{ICC})}, \quad b = 1 + \frac{k \times \widehat{ICC} \times (n-1)}{n \times (1 - \widehat{ICC})}$
- *ICC from two-way random effects (mean of multiple measurements)*		
$ICC_k = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_O^2 + \sigma_{I+A}^2}{k}}$	$\frac{n \times [MS_G - F_{0.975}(n-1, v) \times MS_E]}{F_{0.975}(n-1, v) \times (MS_O - MS_E) + n \times MS_G}$ where v is defined as above	$\frac{n \times (F_{0.975}(v, n-1) \times MS_G - MS_E)}{MS_O - MS_E + n \times F_{0.975}(v, n-1) \times MS_G}$ where v is defined as above

Calculation of the intra class correlation for log-normal data

In contrast to the standard error of measurement, the intra class correlation is a unit-free parameter, expressed as a ratio of variances. Thus, it can still be calculated and interpreted in a similar fashion when the estimation of the variance components are based on log-transformed measurements [56].

2.3.3.4. The coefficient of variation (CV)

The coefficient of variation (CV) is commonly reported in both laboratory and non-laboratory studies. This parameter allows the variability in the measurements produced within individuals (i.e., the measurement error) to be interpreted in relation to the grand mean of the measurements, with higher values indicating a higher variability in the measurements [16, 17]. In mathematical notation, this is expressed as

$$CV = \frac{SEM}{\mu}, \tag{2.19}$$

where SEM is the standard error of measurement, and μ is the grand mean of the measurements. For multiple measurements produced within an individual it can then be stated that, assuming underlying normality, there is a 68% chance of the difference between a produced measurement and the true value of the individual lying within $CV\%$ of the grand mean [16]. A value of ≤10% is commonly used as an indicator of acceptable within-individual variability [16, 57], while Aronhime et al [58] provided the following categorisation for the interpretation of the coefficient of variation: excellent reproducibility when CV≤10%; good reproducibility when CV between 10% and 20%; acceptable reproducibility when CV between 20% and 30%; poor reproducibility when CV>30%.

95% confidence intervals for the coefficient of variation

McKay proposed a method for obtaining an exact confidence interval for the coefficient of variation [59]. Under this method, the lower and upper 95% confidence bounds are calculated by solving the following equations for CV_L and CV_U

$$\left[0.025 = F_{NCT}\left(N-1, \frac{\sqrt{N}}{CV_L}\right)\left(\frac{\sqrt{N}}{CV}\right),\ \ 0.975 = F_{NCT}\left(N-1, \frac{N}{CV_U}\right)\left(\frac{\sqrt{N}}{CV}\right)\right], \tag{2.20}$$

where N is the total number of observations, and $F_{NCT}(p, q)$ represents a non-central T distribution with non-centrality parameter q and p degrees of freedom.

Calculation of the coefficient of variation for log-normal data

Like the intra class correlation, the coefficient of variation is a unit-free parameter. However, this parameter has a meaningful interpretation only when the mean is derived from data-points that are higher than or equal to zero. Thus, when the log-transformation is applied, the calculation based on equation 2.19 is no more applicable, as negative values may arise in the data. Bland and Altman proposed a method for calculating the coefficient of variation when the data are log-normally distributed [60]. Under Bland and Altman, the coefficient of variation is calculated as

$$CV = e^{SEM^{log}} - 1, \tag{2.21}$$

where SEM^{log} is the standard error of measurement estimated on the logarithmic scale (after the log-transformation is applied), and $e^{SEM^{log}}$ is the anti-log. An alternative approach, which is more frequently used in the laboratory setting, is provided by Cole [61]. Using this method, the exact relationship between CV and SEM^{log} is given by

$$CV = \sqrt{e^{(SEM^{log})^2} - 1} \tag{2.22}$$

The two methods are expected to produce very similar estimates, particularly for small values of SEM^{log}.

2.3.3.5. The index of individuality (II)

The Index of individuality is a parameter used in studies of laboratory tests. It is calculated as the ratio of the variability at the random analytical and within-individual biological level to the true variability between the individuals, with lower values indicating a higher between-individual variability in relation to the measurement error [17, 27, 62]. In mathematical notation, the parameter is expressed as

$$II = \frac{\sqrt{\sigma_A^2 + \sigma_I^2}}{\sigma_G}, \tag{2.23}$$

where $\sigma_A, \sigma_I, \sigma_G$ are the estimated standard deviations for the random analytical, within-individual biological, and between-individual biological components of variability, respectively. Although recommended against, the calculation is often simplified to $II = \frac{\sigma_I}{\sigma_G}$, under the assumption that $\sigma_A \ll \sigma_I$ [17, 63]. Harris proposed the use of the index of individuality for evaluating the utility of specifying population-based reference intervals for making decisions on whether a change in the health status of the individual has occurred [62]. High index values (≥1.4) indicate that the measurements produced within individuals can be compared usefully with a reference interval, as the measurements will occupy most of the interval, or even fall outside the interval if a true change occurs. Iron is an example of a analyte known to have a high index of individuality [17]. For low index values (≤0.6) however, the usefulness of a reference interval for interpreting test results is limited, as the range of the within-individual measurements will only cover a small part of the interval.

Common analytes with a low index of individuality include serum alanine aminotransferase, creatine kinase, magnesium, uric acid, and serum creatinine [17].

95% confidence intervals for the index of individuality

No methods have been proposed for calculating a confidence interval for the index of individuality.

Calculation of the index of individuality for log-normal data

The index of individuality is a unit-free parameter, expressed as a ratio of variances. Thus, it can still be calculated and interpreted in a similar fashion when the estimation of the variance components are based on log-transformed measurements.

2.3.3.6. The reference change value (RCV)

For biomarkers with a low index of individuality (≤0.6), Harris and Yasaka proposed the use of the reference change value for evaluating changes in the health status of an individual [64]. This parameter may be used to calculate the minimal difference between two successive measurements of an individual that needs to be exceeded, in order to state that a true change in the health status has taken place. The reference change value is calculated as

$$RCV = \sqrt{2} \times Z \times CV_{I+A}, \tag{2.24}$$

where Z represents a *Z*-score selected from the normal distribution, $CV_{I+A} = \frac{\sqrt{\sigma_A^2+\sigma_I^2}}{\mu}$, $\sqrt{\sigma_A^2 + \sigma_I^2}$ is the total within-individual error, and μ is the grand mean of the measurements. Fraser suggests using 1.96 and 2.58 as the *Z*-scores for a significant and a highly significant change, respectively. If for example the first PSA (prostate-specific antigen) measurement of an individual is 7.3μg/L, and

the reference change value based on *Z*=1.96 is estimated to be 51%, then repeated measurements above or below 7.3 × (51/100) = 3.7µg/L are interpreted as a significant change. This parameter is akin to the smallest detectable change but with the measurement error expressed in relation to the grand mean of the measurements.

95% confidence intervals for the reference change value

When desired, confidence intervals for the reference change value can be constructed using

$$[\sqrt{2} \times Z \times CV_{I+A}^{L},\ \sqrt{2} \times Z \times CV_{I+A}^{U}\ ], \tag{2.25}$$

where $CV_{I+A}^{L}, CV_{I+A}^{U}$ are the lower and upper confidence bounds of CV_{I+A}, which can be obtained as described under 2.3.3.4.

Calculation of the reference change value for log-normal data

Fokkema et al proposed methods for obtaining an asymmetrical interval for the RCV [65], with the lower and upper bounds calculated as

$$[RCV_L = \left(-\sqrt{2} \times Z \times \tau - 1\right) \times 100, RCV_U = \left(\sqrt{2} \times Z \times \tau - 1\right) \times 100], \tag{2.26}$$

where $\tau = \sqrt{(\log(CV_{I+A}^{2} + 1)}$, and RCV_L and RCV_U represent the lower and upper bound, respectively. In order to state that a true change has taken place, the ratio of two consecutive measurements produced from the same individual must fall outside this interval.

2.3.3.7. The limits of agreement (LoA)

An alternative non-regression based approach for measuring the variability between two measurements produced within the individuals includes the limits of agreement, proposed by Bland and Altman [15, 16, 47]. This method is most commonly encountered in non-laboratory studies, and accounts for both potential random and systematic variability between two measurements produced within each individual. The limits of agreement are calculated as the interval of 1.96 times the standard deviation of the differences between two measurements produced from the same individual, either side of the mean difference [15, 47]. In mathematical notation this is expressed as

$$LoA = \bar{d} \pm 1.96 \times SD_{\bar{d}}, \tag{2.27}$$

where $\bar{d} = \sum_{i=1}^{n} \frac{d_i}{n}$ represents the systematic differences between two repeated measurements produced within the individuals, y_{i1} and y_{i2}, d_i indicates the difference between the two measurements ($d_i = y_{i1} - y_{i2}$), and $SD_{\bar{d}} = \sqrt{\sum_{i=1}^{n} \frac{(d_i - \bar{d})^2}{n-1}}$ represents the random variability occurring between the measurements. It can then be stated with 95% confidence that differences between two within-individual measurements falling within the produced interval are only due to measurement error, and not due to a true change in the performance of the individual [15, 47]. This method requires the following two assumptions to be met: (i) the differences between the two measurements produced within-individuals should be normally distributed, and (ii) there should be no evidence of heteroscedasticity in the data. In contrast to any other method described so far, this method can only be used when a pair of measurements is available from each individual.

95% confidence intervals for the limits of agreement

Bland and Altman recommend that the produced interval is presented along with the lower 95% confidence bound of the lower limit of agreement, and the upper 95% confidence bound of the upper limit of agreement [47]. These were calculated as

$$95\%CI - LoA_{U/L} = \left(\bar{d} \pm 1.96 \times SD_{\bar{d}}\right) \pm t\left(n_{pairs} - 1, 0.025\right)\sqrt{\frac{3 \times SD_{\bar{d}}{}^2}{n_{pairs}}}, \tag{2.28}$$

where n_{pairs} is the number of available pairs of measurements, and $t(n_{pairs} - 1, 0.025)$ is the critical value of the Student's t distribution with $n_{pairs} - 1$ degrees of freedom and 2.5% significance level.

Calculation of the limits of agreement for log-normal data

Euser et al proposed methods for calculating the limits of agreement when the measurements are log-transformed [56]. In this case, the limits of agreement are expressed as a function of the mean of a randomly selected pair of measurements produced from the same patient. If Y_1 and Y_2 are two measurements produced from the same individual, then the difference between the two measurements $(Y_1 - Y_2)$ lies within

$$LoA = \pm\frac{2\bar{Y}(10^a - 1)}{10^a + 1}, \tag{2.29}$$

where $\bar{Y} = \frac{Y_1 + Y_2}{2}$, and $a = 1.96 \times SD_{\bar{d}}$.

2.3.3.8. The Pearson correlation coefficient (r)

The Pearson correlation coefficient is often reported in studies of physiologic tests [15, 66]. It is calculated as

$$r = \frac{Cov(y_1, y_2)}{\sigma_{y_1}\sigma_{y_2}} = \frac{\sum_{i=1}^{n}(y_{i1} - \overline{y_1})(y_{i2} - \overline{y_2})}{\sqrt{\sum_{i=1}^{n}(y_{i1} - \overline{y_1})^2 \sum_{i=1}^{n}(y_{i2} - \overline{y_2})^2}}, \quad (2.30)$$

where (y_{i1}, y_{i2}) is a pair of measurements produced from the i_{th} patient $(i = 1, \dots, n)$, $\overline{y_1}$ and $\overline{y_2}$ are the estimated mean values for y_1 and y_2, σ_{y_1} and σ_{y_2} are the corresponding standard deviations, and $Cov(y_1, y_2)$ is the covariance.

This parameter is used for assessing the reliability of the produced measurements, and can be interpreted in a similar fashion to the intra class correlation coefficient given that potential variability arising from systematic differences between y_1 and y_2, is absent. If not absent, this parameter will provide an inadequate estimate of reliability. The use of the Pearson correlation coefficient additionally requires the following assumptions to be met: (i) the measurements produced at the initial and repeated testing sessions should both be normally distributed, (ii) there is a linear relationship between the measurements produced at each session (iii) no extreme outliers are observed. Similar to the limits of agreement, this parameter can only be used when a pair of measurements is available from each individual.

95% confidence intervals for the Pearson correlation coefficient

The most popular method used for constructing a confidence interval for the Pearson correlation coefficient is that based on the *Z*-transformation, which was originally proposed by Fisher in 1921 [67]. The estimate produced from equation 2.30 is transformed to a *Z*-statistic using

$$Z = 0.5 \ln\left(\frac{1+r}{1-r}\right) \quad (2.31)$$

Under this transformation, the sampling distribution of the Pearson correlation coefficient approximates normality. Based on this assumption, a 95% confidence interval for the *Z*-statistic can be constructed as

$$Z_{U/L} = Z \pm 1.96\sqrt{\frac{1}{(n-3)}}, \tag{2.32}$$

where $Z_{U/L}$ is the upper/lower 95% confidence bound for the Fisher's *Z* statistic, and $\sqrt{\frac{1}{(n-3)}}$ is the corresponding standard error. The produced estimate and 95% confidence intervals are then reverted to the r metric using

$$r = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1} \tag{2.33}$$

Calculation of the Pearson correlation coefficient for log-normal data

When the log-transformation is applied, the calculation based on equation 2.30 is still valid, given that the aforementioned requirements for using the Pearson correlation coefficient as a reliability parameter are still met. However, when skewness in the measurements is observed, an alternative approach to log-transforming the measurements, which is recommended by several authors [68, 69], is the use of the Spearman correlation coefficient. This statistic is considered as the non-parametric version of the Pearson correlation coefficient, and the calculation is based on the ranks of the measurements produced at each testing session (rather than the actual measurements). If (y_{i1}, y_{i2}) is a pair of measurements produced from the i_{th} patient $(i = 1, \dots, n)$, the Spearman correlation is calculated as

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n}(rank[y_{i1}] - rank[y_{i2}])^2}{n(n^2 - 1)}, \tag{2.34}$$

where $rank[y_{i1}]$ and $rank[y_{i2}]$ are the ranks of y_{i1} and y_{i2}, respectively.

2.3.3.9. The Kappa coefficient

The Kappa coefficient, which was first introduced by Jacob Cohen in 1960 [70], is a statistical parameter used for evaluating variability between multiple observers (inter-observer variability) or

within the same observer (intra-observer variability), when assessing the same individual multiple times [15]. This parameter is mostly reported in studies of imaging tests. In contrast to all the previous methods described so far, where the produced response is measured on a continuous scale, the use of this parameter requires the data to be binary (e.g., condition absent/present) or ordinal (e.g., condition absent/mild/moderate/severe). The advantage of the Kappa coefficient over the standard way of expressing agreement as percentage is the ability to account for any agreement which is expected by chance. Guidelines for the interpretation of the Kappa coefficient (presented in Table 2.6) have been provided by Landis and Koch [71] and Fleiss [72].

Table 2.6. Interpretation of the Kappa value based on Landis and Koch [71], and Fleiss [72].

Landis and Koch [71]			Fleiss [72]	
Kappa between 0 and 0.20	slight		Kappa less than 0.40	poor
Kappa between 0.21 and 0.40	fair			
Kappa between 0.41 and 0.60	moderate		Kappa between 0.40 and 0.75	fair to good
Kappa between 0.61 and 0.80	substantial			
Kappa between 0.81 and 1	excellent		Kappa higher than 0.75	excellent

Kappa coefficient for binary data

The Kappa coefficient is calculated as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{2.35}$$

where p_o is the proportion of cases where agreement between two observers (or within the same observer) is achieved, and p_e the proportion of cases where the agreement is expected by chance [15]. These are calculated as

$$p_o = \frac{a + d}{a + b + c + d}, \tag{2.36}$$

and

$$p_e = \frac{(a + b) \times (a + c) \times (c + d) \times (b + d)}{(a + b + c + d)^4}, \tag{2.37}$$

where a, b, c, d are the values within Table 2.7.

Table 2.7. Example of two observers assessing presence or absent of a condition.

First observer	**Second observer**		
	Condition present	Condition absent	Total
Condition present	a	b	$a + b$
Condition absent	c	d	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

Kappa coefficient for ordinal data

An extension of this method, called the weighted Kappa, can be employed when the produced response is measured on an ordinal scale [15, 73]. The rationale of weighted Kappa is that a misclassification between two adjacent categories is of less concern compared to categories that are

more distant. Thus, the latter should be penalized more heavily. The weighted Kappa is again estimated using equation 2.35, but this time the calculation of p_o and p_e are extended to

$$p_o = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_{ij}, \tag{2.38}$$

and

$$p_e = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_i p_j, \tag{2.39}$$

where p_{ij} is the proportion of individuals assigned to the i_{th} category by the first observer and the j_{th} category by the second observer, $p_i = \sum_{1\leq j \leq n} p_{ij}$ is the proportion of individuals assigned to the i_{th} category by the first observer, and $p_j = \sum_{1\leq i \leq n} p_{ij}$ is the proportion of individuals assigned to the j_{th} category by the second observer. Using the method proposed by Cicchetti [15, 73], a weight to the (i, j) cell of the $n \times n$ table can be assigned as follows

$$w_{ij} = 1 - \frac{|i-j|}{n} \tag{2.40}$$

95% confidence intervals for the Kappa coefficient

A confidence interval using the normal approximation can then be obtained as

$$\kappa \pm 1.96 \times SE_\kappa, \tag{2.41}$$

where SE_K is the standard error of the Kappa coefficient. For binary outcomes, the standard error is calculated as

$$SE_\kappa = \sqrt{\frac{p_o(p_o - p_e)}{n(1 - p_e)^2}} \tag{2.42}$$

For the weighted Kappa, the calculation is extended to

$$SE_{\kappa_W} = \frac{1}{(1-p_e)\sqrt{N}} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (w_{ij} - \sum_{i=1}^{n} w_{ij} p_i - \sum_{j=1}^{n} w_{ij} p_j)^2 - p_e^2} \quad (2.43)$$

2.3.4. Additional methods for constructing confidence intervals

The use of confidence intervals has been recommended by many researchers of the field, as it quantifies the uncertainly around a parameter estimate [26, 47, 74]. Methods available for constructing a confidence interval for each parameter were presented in section 2.3.3. For some parameters however, the construction of confidence intervals involves solving complex equations, while for others, methods for constructing a confidence interval may perform poorly under specific scenarios, or even not exist. Two alternative methods for obtaining confidence intervals include the multivariate delta and bootstrap methods. Both methods are simple to be implemented given the modern computing power, and available in most statistical software packages.

2.3.4.1. The multivariate delta method

The multivariate delta method is a statistical technique for deriving a standard error for a function of parameters, whose estimators follow an asymptotically normal distribution. The method was first described by Doob in 1935 [75]. Let $\theta = (\theta_0, \theta_1)$ be a random vector of two statistical parameters with corresponding variances $\sigma^2_{\theta_0}$ and $\sigma^2_{\theta_1}$, and $\hat{\theta} = (\widehat{\theta_0}, \widehat{\theta_1})$ be an unbiased estimate of each element of θ. If

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \sim N(0, \sigma^2_{\theta}) \quad (2.44)$$

as the sample size n increases, then for any given function $g(\theta) = g(\theta_0, \theta_1)$ with continuous first partial derivatives, it follows that

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \sim N(0, V_g) \tag{2.45}$$

where $V_g = \frac{g(\theta)}{\partial\theta_0} \times V_\theta \times \frac{g(\theta)}{\partial\theta_1}$, and $V_\theta = \begin{pmatrix} \sigma^2_{\theta_0} & \sigma_{\theta_0\theta_1} \\ \sigma_{\theta_0\theta_1} & \sigma^2_{\theta_1} \end{pmatrix}$. A 95% confidence interval is then constructed as

$$g(\hat{\theta}) \pm 1.96 \times \sqrt{V_g} \tag{2.46}$$

, where $\sqrt{V_g}$ is the delta method-produced standard error. Similarly, this method can also be applied for deriving the standard error for a function of three or more parameters (e.g., ICC produced from a two-way random effects model). With respect to the parameters presented in section 2.3.3, Curto and Pinto [76] used this method to derive a standard error for the coefficient of variation $\left(SE_{\widehat{CV}} = \sqrt{\frac{\left(\widehat{CV}^4 + \frac{\widehat{CV}^2}{2}\right)}{n}} \right)$. Another application of this method includes the derivation of a standard error for the Pearson correlation coefficient $\left(SE_{\hat{r}} = \frac{(1-\hat{r}^2)^2}{n-1} \right)$ [77].

2.3.4.2. The bootstrap technique

The concept of bootstrap was introduced by Efron in 1979 [78]. Under this method, multiple random samples of the same size as the original sample, are repeatedly drawn with replacement from the data, and in turn analysed in order to obtain a sampling distribution for the statistical parameter(s) of interest. When the produced distribution is used for constructing confidence intervals, at least 1000 bootstrap samples are required [79]. Several approaches can then be used for constructing a bootstrap-based confidence interval [78, 80]. These include the normal approximation, the percentile, and the bias-corrected and accelerated methods.

Confidence intervals using a normal approximation

When using this approach, a 95% confidence interval is obtained as

$$\hat{\theta} \pm 1.96 \times \sigma_{\hat{\theta}}, \tag{2.47}$$

where $\sigma_{\hat{\theta}}$ denotes the standard deviation of the produced bootstrap estimates. Several authors recommend against constructing normally approximated confidence intervals, as this approach assumes that there is no bias in the produced bootstrap estimates (i.e. the median of the obtained sampling distribution equals the original sample estimate), an assumption which is often violated, and only makes use of $\sigma_{\hat{\theta}}$ without exploiting the sampling distribution of the bootstrap estimates [80].

Percentile confidence intervals

Under this approach, the lower and upper 95% confidence bounds are equal to the 2.5% and 97.5% percentiles of the obtained sampling distribution, respectively. Alike the normal approximation, this approach assumes that the bootstrap estimates are unbiased.

Bias-corrected confidence intervals

As the name implies, the bias-corrected percentile approach is recommended when bias is present (i.e., the original sample estimate does not lie at the 50th percentile of the sampling distribution) [79, 80]. Under this approach, a bias-correcting constant is calculated as

$$Z_0 = \Phi^{-1}\left(\frac{\sum_{1 \le i \le n_B} \mathrm{N}(\hat{\theta}_i \le \theta)}{n_B}\right), \tag{2.48}$$

where Φ denotes the cumulative distribution of the standard normal function, $\hat{\theta}_i$ denotes the estimate obtained from the i_{th} bootstrap sample, θ denotes the estimate of the original sample, n_B denotes the total number of bootstrap samples, and $\mathrm{N}(\hat{\theta}_i \le \theta)$ is a binary indicator of whether the i_{th} bootstrap estimate was higher than the original sample estimate $(\mathrm{N}(\hat{\theta}_i \le \theta) = 0)$ or not $(\mathrm{N}(\hat{\theta}_i \le \theta) = 1)$. The lower and upper 95% confidence bounds are then equal to the $[\Phi(-1.96 + 2Z_0) \times 100]$ and $[\Phi(1.96 + 2Z_0) \times 100]$ percentiles of the obtained sampling distribution, respectively.

Bias-corrected and accelerated confidence intervals

In addition to the correction of any potential bias, this approach accounts for any potential skewness in the distribution of bootstrap estimates [80]. The construction of the 95% confidence intervals is based on the assumption that a monotonic increasing function f exists, such that

$$[f(\hat{\theta}) - f(\theta)] \sim N(-Z_0 \times \tau_f, \tau_f^2) \tag{2.49}$$

, where Z_0 is the standard normal deviate (defined as above), $\tau_f = 1 + \alpha \times f(\hat{\theta})$ denotes the standard error of $f(\hat{\theta})$, and α is the acceleration parameter. A 95% confidence interval for $\hat{\theta}$ is constructed by applying the inverse function f^{-1} to the 95% confidence intervals of $f(\hat{\theta})$. This is mathematically expressed as

$$95\%\ CI_{\hat{\theta}} = f^{-1}[(f(\hat{\theta}) + Z_0 \times \tau_f) \pm 1.96 \times \tau_f] \tag{2.50}$$

2.4. Reporting of primary studies examining the measurement error of biomarkers

Transparent reporting enables the results published in manuscripts to be interpreted appropriately, replicated by other researchers, and in turn used for clinical decision and policy making [81-84]. As such, the presence of reporting guidelines play a critical role in medical research. With respect to the measurement error, Bartlett et al [85] developed guidelines for appraising the existing and future publications of studies of laboratory-based tests, while for guidelines for reporting studies of non-laboratory tests were developed by Kottner et al [36]. These guidelines are presented in Table 2.8.

Table 2.8. Reporting guidelines provided by Bartlett et al [85] and Kottner et al [36].

Bartlett et al [85]		Kottner et al [36]	
Title	This should indicate that the content relates to a study of biological variability, the subject of the study, the sample matrix, and the studied population.		
Abstract	This section should include (as a minimum) the headline biological variability data, the number and demographic characteristics of the study participants, a clear identification of the target analyte/measurand, the statistical methods used, and the geographical location of the study.	**Title & abstract**	This should include a clear description of the type of variability examined (e.g., inter/intra-observer variability).
Introduction	This section should include a clear description of the context and aims of the study, and citation of any previous studies of biological variability of the analyte of interest.	**Introduction**	This should provide information on the testing equipment, the population investigated, the participating observers, and the rationale for the study.
Methods	This section should provide information on the analyte/measurant of interest, the study population, the length of the study, and the collection/analysis/storage conditions of the samples.	**Methods**	This section should provide justification of the chosen sample size, and a clear description of the measurement process and statistical analysis.
Data analysis	This section should include detecting and excluding outliers, assessing the heterogeneity of variance, and		

	using appropriate statistical methods (which should be clearly described).			
Results	This section should include a clear presentation of the biological variability estimates and corresponding confidence intervals, using unified terms and symbols.		**Results**	This should provide information on the number of individuals, observers, and measurements performed within individuals, a clear description on the characteristics of the individuals and observers, and a clear presentation of the variability estimates and the corresponding statistical uncertainty.
Discussion	This section should include a description of the strengths and limitations of the study, and a focus on factors that may potentially affect the transportability of the findings to other settings.		**Discussion**	This should provide details on how the results can be applied in practice.

2.5. General statistical methods used for the meta-analysis of data reported across primary studies

Meta-analysis involves using statistical methods to synthesize quantitative evidence from primary studies, so that an overall estimate is obtained based on a whole body of research [31, 86]. Although well established for statistical parameters used in other areas of medical research (e.g., prognostic or diagnostic research, effectiveness of new interventions) [87-89], methods for performing a meta-analysis are less developed for parameters of the measurement error of biomarkers. As such, the thesis aims to explore the state-of-the-art in the meta-analytic methods used for parameters of measurement error (see Chapter 4). Before investigating methods specific to parameters expressing the measurement error of biomarkers, this chapter introduces known methods for meta-analysis in general, which are applied across all areas of medical research.

2.5.1. Meta-analysis using aggregate study-level data

The traditional approach for performing a meta-analysis involves combining summary estimates of the statistical parameter of interest, which are provided from the studies, while accounting for how precisely the parameter is estimated within each study [31]. Two statistical models are available for this purpose. These include the fixed-effect model and the random-effects model [90]. Both models use similar sets of calculations to compute a pooled estimate across studies (see sections 2.5.1.1 and 2.5.1.2) and may often produce a similar result. However, the two models cannot be used interchangeably, as they hold different assumptions about the reported estimates.

2.5.1.1. The fixed-effect model

The simplest model that is commonly used in meta-analysis is the fixed-effect model, which is also often referred to as the common-effects model [91]. Let $y_1, y_2, \dots, y_k$ be estimates of a particular parameter, provided by a set of k independent studies. The fixed-effect model assumes that the

studies are estimating a common true effect, μ, and that each y_i $(i = 1,2, \dots, k)$ can only differ from μ due to the within-study sampling error variance (i.e., variability created from the patients sampled in each individual study). In model notation, this can be expressed as

$$y_i = \mu + \varepsilon_i, \tag{2.51}$$

and $\varepsilon_i \sim N(0,\ \sigma_i^2)$ represents the estimation error due to the within-study sampling variability, where σ_i^2 is the variance estimate of the i_{th} study. A weighted average estimate is then calculated based on the method of *inverse variance weights.* This method ensures that a study reporting a more precise estimate is assigned a greater weight in the meta-analysis. The calculation proceed as follows

$$\bar{y} = \frac{\sum_{1 \le i \le k} w_i y_i}{\sum_{1 \le i \le k} w_i}, \tag{2.52}$$

where w_i represents the wight assigned to the i_{th} study, and is equal to

$$w_i = \frac{1}{\sigma_i^2} \tag{2.53}$$

The variance of $\bar{y}$ is computed as

$$Var_{\bar{y}} = \frac{1}{\sum_{1 \le i \le k} w_i}, \tag{2.54}$$

with a 95% confidence interval being constructed as

$$95\%\ CI_{\bar{y}} = \bar{y}\ \pm 1.96 \times \sqrt{Var_{\bar{y}}} \tag{2.55}$$

2.5.1.2. The random-effects model

In contrast to the fixed-effect model, this model allows the observed estimates across studies to vary due to within-study error variance, as well as any potential between-study heterogeneity (i.e., real differences across the study characteristics) [92]. Under this model, the observed estimate in the i_{th}

study ($i = 1, \dots, k$) is sampled from a normally distributed population with mean μ_i and variance σ_i^2. The mean of each study population is in turn sampled from a larger population, normally distributed with mean μ and variance τ^2, where τ^2 is the variance attributed to differences in the means of the populations the studies were sampled. In model notation, this can be expressed as

$$y_i = \mu_i + \varepsilon_i = (\mu + \delta_i) + \varepsilon_i \tag{2.56}$$

were $Y_i \sim N(0,\ \sigma_i^2 + \tau^2)$ represents the observed estimate in the i_{th} study, μ represents the overall population mean, $\delta_i \sim N(0,\ \tau^2)$ represents the difference between the overall population mean (μ) and the mean of the population the i_{th} study was sampled (μ_i), and $\varepsilon_i \sim N(0,\ \sigma_i^2)$ represents the estimation error due to within-study sampling variability. If $\tau^2 = 0$, the model reduces to a fixed-effect.

A weighted average estimate along with its corresponding variance and 95% confidence intervals are computed in a similar fashion to a fixed-effect analysis (see equations 2.52, 2.54, and 2.55). However, the calculation of the weight assigned to each study is this time extended to

$$w_i = \frac{1}{\sigma_i^2 + \tau^2} \tag{2.57}$$

2.5.1.3. Statistical methods used to explore between-study heterogeneity

The Cochran's Q is the most commonly used test for examining whether significant between-study heterogeneity is present [93]. Under the null hypothesis that the underlying effect μ does not differ across studies, a p-value is obtained by comparing the produced Q statistic to a chi-square distribution with $k - 1$ degrees of freedom, with k representing the number of studies. The value of the Q statistic is computed as

$$Q = \sum_{1 \le i \le k} w_i \, (y_i - \bar{y}\,)^2 \tag{2.58}$$

However, the test often fails to detect significant heterogeneity, particularly when the number of studies included in the meta-analysis is small [94]. As such, the level of statistical significance is often

set at 10%, with a p-value<0.1 being considered an indicative of substantial between-study heterogeneity [95, 96].

An alternative approach developed by Higgins et al includes the use of the I-squared statistic [96]. This quantity describes the percentage of total variation in the meta-analysis that is due to between-study heterogeneity. It is calculated as

$$I^2 = \frac{Q - (k-1)}{Q} \times 100 \tag{2.59}$$

and ranges from 0% to 100%, with larger values indicating increasing heterogeneity. See Table 2.9 for a guide to interpretation of the produced value based on the Cochrane Handbook for Systematic Reviews of Interventions [97].

Table 2.9. Interpretation of I-squared value based on the Cochrane Handbook.

Value of I-squared statistic	Interpretation
Less than 40%	Heterogeneity might not be important
Between 30% and 60%	Heterogeneity may be moderate
Between 50% and 90%	Heterogeneity may be substantial
Between 75% and 100%	Heterogeneity is considerable

However, Borenstein et al [98] recommend against using the produced I-squared value for assessing between-study heterogeneity, as this statistic does not provide any information on how the individual study-level means (μ_i) vary. Furthermore, Migliavaca et al [99] state that a high (or low) I-squared value is not always synonymous with the presence (or absence) of significant between-study heterogeneity, as the calculation may be affected by several factors, such as the number of studies included in the meta-analysis or the type of outcome being pooled (e.g., meta-analyses of proportional data often yield I-squared values>95% [99]). Instead of the I-squared statistic, many researchers advocate the use of the prediction interval for obtaining a range of values for μ_i that is

expected across different study populations [98-100]. The lower and upper bounds of the prediction interval can be calculated as

$$PI_{low} = \bar{y} - t_{k-2} \times \sqrt{Var_{\bar{y}} + \tau^2}, \tag{2.60}$$

and

$$PI_{upp} = \bar{y} + t_{k-2} \times \sqrt{Var_{\bar{y}} + \tau^2}, \tag{2.61}$$

where $\bar{y}$ and $Var_{\bar{y}}$ are the weighted average estimate and its corresponding variance, τ^2 is the estimated between-study variance, and t_{k-2} is the 97.5th percentile of the t distribution with $k-2$ degrees of freedom.

It can then be stated that in 95% of all populations, the value of the meta-analytic parameter of interest is expected to lie within (PI_{low}, PI_{upp}). The use of the 95% prediction intervals is applicable only under a random-effects meta-analysis, as a fixed-effect meta-analysis assumes that the included study-level estimates are sampled from the same population.

2.5.1.4. Estimation of the between-study variance (τ^2)

The most popular method available for estimating the between-study variance is the method of moments, proposed by DerSimonian and Laird [89]. Under this method, the between-study variance is estimated via

$$\tau^2 = \frac{Q - (k-1)}{C}, \tag{2.62}$$

where C is a scaling factor accounting for the fact that Q is a weighted sum of squares [90], computed as

$$C = \sum_{1 \le i \le k} w_i - \frac{\sum_{1 \le i \le k} w_i^2}{\sum_{1 \le i \le k} w_i} \tag{2.63}$$

Alternative methods for estimating the between-study variance are also available and based on maximum likelihood estimation, with evidence suggesting that these methods are able to provide a more precise estimate, particularly when the number of studies is small [101]. However, likelihood-based methods are more computationally intensive, while Borenstein et al state that no method can produce a very precise estimate when the number of studies is small [90]. For meta-analyses including a small number of studies (i.e., <5), guidelines provided in the Cochran Handbook recommend a different approach based on the Bayesian framework, where prior distributions are used for the extent of between-study variation [102].

2.5.1.5. Which model should be used?

Borenstein et al [90] state that a fixed-effect model can be employed under the following two conditions:

- the characteristics of the studies are functionally identical. For studies examining the reliability and measurement error of biomarkers, these may for example include the demographics of the individuals, the experience of the clinicians, or a different test-retest interval being adopted.
- the produced summary estimate reflects the included studies only, and does not aim across all possible studies possessing similar (but not identical) characteristics.

By contrast, a random-effects model should be preferred if the aim of the meta-analysis is to extrapolate the result to a wider population (which is most often the case), or if the characteristics differ across studies (which is likely to happen when different researchers operate independently). Furthermore, the decision on whether to use a fixed-effect or a random-effects approach should be purely driven by the two aforementioned conditions rather than the use of a statistical test for heterogeneity, as these tests often suffer from low power and may in some cases be misleading [96, 103].

2.5.2. Meta-analysis using individual participant data

A more recent approach, which has become increasingly popular, is to perform a meta-analysis of individual participant data [86, 104]. Rather than obtaining a parameter estimate and its corresponding uncertainty (e.g., standard error or 95% confidence interval) from each identified study, this approach involves collecting raw data for each participant recruited within each study. Once all data has been collected, the two possible ways for producing a summary estimate include:

- The *one-step approach*, where the raw data collected from the studies are treated as a single large data set, and analysed simultaneously. When using this approach, it is important to account for the clustering of participants within the studies. This involves treating the study effect as fixed (if only the included studies are or interest) or random (if the aim is to generalise the produced estimate beyond the included studies) [86].
- The *two-step approach*, where each data set obtained for each study is first analysed independently. The produced estimates are then pooled in a similar fashion to an aggregate study-level meta-analysis (described under section 2.5.1).

Compared to the traditional method of pooling study-level estimates, the use of this approach has several advantages. These include the ability to explore the data in detail, the use of consistent eligibility criteria and statistical methods across the studies, the inclusion of more patients or patients with a longer follow-up than that used in the original study publication, and the use of information from studies where the aggregate data are not available or poorly reported [86]. However, performing a meta-analysis of individual participant data may not be worth in case the required aggregate study-level estimates can be fully obtained from the published primary studies or the corresponding authors, as collecting the raw data of each individuals may require considerable time and resources.

2.6. Discussion

This chapter provides an overview of the current guidelines for the design and statistical analysis of primary studies examining the reliability and measurement error of biomarkers. However, these guidelines have several flaws, and should be further developed. When designing primary studies examining the reliability and measurement error of biomarkers, different views have been expressed regarding the population that should be recruited in the study. Current guidelines for laboratory-based studies suggest that the population should be limited to healthy individuals only, as the focus of such studies is to examine the biological sources of variability, rather than any pathological [27]. However, for variability estimates to be of use in practice, the recruited individuals should reflect the population that is of interest. If for example a test is intended for individuals with a particular disease, then there is no use testing healthy individuals, as the variability of a test may change between healthy and diseased populations. When recruiting diseased individuals however, an obvious concern for researchers designing such studies should be the potential progression in between the repeated measurements. Therefore, an appropriate test-retest interval should be carefully chosen by the researchers, so that any potential bias in the results due to disease progression is kept to a minimum. The appropriate interval will be specific to the disease of interest and will need to be considered in view of the likely progression of the disease.

Furthermore, there is limited guidance regarding the number of individuals required for designing primary studies examining the measurement error of biomarkers. Fraser and Harris [17, 26] and Braga and Panteghini [27] state that this decision should be a trade-off between having a high number of individuals (and consequently samples) that will allow the different sources of variability to be estimated more precisely, and a smaller number that will enable the samples to be handled appropriately and analysed under the right conditions. However, this trade-off is not necessarily inevitable, given that a study protocol is clearly specified and strictly followed by the operators. Giraudeau and Mary provide a formula for estimating the number of individuals required, based on a number of measurements per individual, a pre-specified estimate for the intra class correlation (ICC),

and a pre-specified width for the 95% confidence interval of the ICC estimate [43]. However, the formula is limited to ICC's produced from a one-way effects model (i.e., assuming no inter/intra-observer variability), and formulas appropriate for ICC's based on two-way effects models are currently lacking. Moreover, whilst the ICC expresses how reliable a test is (i.e., how large the between-individual variability is in relation to the measurement error), it does not provide any direct information regarding the absolute deviation of repeated within-individual measurements. Thus, new methods for estimating the required sample size should be developed, additionally accounting for the case were inter/intra-observer variability may be present, and focusing on the compromise between a high reliability and low measurement error.

For the preparation of the data prior to analysis, current guidelines provided for laboratory-based studies suggest removing outlies from the data, and using the log-transformation in case the measurements produced at different variability levels are not normally distributed. These guidelines recommend that the presence of outliers should be examined through the Cochran's test [44] and Dixon-Reed's criterion [45], while the Shapiro-Wilks [46] and Kolmogorov–Smirnov [105] tests should be employed to assess the distribution of the measurements. For outliers, it is known that they often occur in real life, and may indicate difficulties when performing a measurement. Thus, outliers should not be removed from the data, as the removal of outliers may lead to underestimating the "true" variability of a test, and in turn to false conclusions regarding whether a test is fit for use in practice. With respect to log-transforming the measurements prior to analysis, this approach should be used with caution, as published evidence suggests that applying this transformation does not always lead to a better approximation of normality [106]. Furthermore, statistical tests for assessing whether within-individual measurements are free from outliers or deviate from normality should also be used with caution (if used at all). This is because primary studies of test variability often collect a limited number of measurements from each recruited individual (e.g., <5). With such low numbers, these tests are likely to produce biased results and in turn lead to misleading conclusions. For example, a statistical test for normality may falsely accept

the hypothesis that measurements are normally distributed [107], allowing researchers to carry on evaluating a laboratory biomarker while they should have stopped (as current guidelines recommend).

With regards to the statistical analysis, several parameters can then be obtained to express the reliability and measurement error of tests producing continuous outcomes, while one additional parameter is available in the literature for assessing inter/intra-observer variability in categorical outcomes (i.e., the Kappa statistic). A key question arises on the interpretation of these parameters. For unit-free parameters of reliability and measurement error, classification tools have been proposed, aiming to help researchers with the interpretation of the produced values. For the intra-class correlation and the Kappa coefficient, two different classification tools are available in the literature (see Tables 2.4 and 2.6). However, the tools do not entirely agree with each other, and the classification of an ICC or Kappa value may differ depending on the tool being chosen by the researcher. For example, an ICC value of 0.80 is considered "good reliability" when using the classification guidelines provided by Koo and Li [53], but "excellent reliability" when the classification guidelines provided by Cicchetti [52]. For the coefficient of variation, no formal guidelines for interpretation have been developed. A cut-off value of ≤10% seems to have been adopted by many researchers as the "working rule" [16, 57]. However, this value has been arbitrarily chosen, and often interpreted in different ways. For example, an estimate of 10% was considered "acceptable" in Stokes [108], but "excellent" in Aronhime et al [58]. Whilst further work is required aiming for the standardisation of the interpretation of these parameters, researchers should also be encouraged to base their interpretations on other factors, rather than any available tools. Such factors may include the severity of the disease of interest (a higher test reliability and lower measurement error is required for life threatening compared to less severe diseases), or the reliability and measurement error of already existing tests (new tests should be more reliable and accompanied with lower measurement error, compared to already existing).

Furthermore, as existing methods for examining the reliability and measurement error are limited to test results expressed as continuous or categorical, alternative methods for count outcomes are also required. Such methods will provide researchers primary evidence of the reliability and measurement error of a count-based test, helping them decide whether the test is fit for use in medical research and practice. Finally, although well-established in other areas of medical research, guidance on how to perform a meta-analysis of parameter estimates of reliability and measurement error is currently lacking. Statistical methods for performing a meta-analysis of data reported in primary studies examining the reliability and measurement error of tests should also be developed, so that conclusions on the reliability and measurement error of tests can be drawn based on a whole body of research.

2.7. Conclusion

Current guidelines for primary studies examining the reliability and measurement error of medical tests have flaws and should be further developed, particularly with respect to the target population, the preparation of data prior to statistical analysis, and the interpretation of statistical parameters expressing the reliability and measurement error of tests. New statistical methods for estimating an adequate number of individuals needed for a primary study are required. These methods should additionally cover the case were inter/intra-observer variability is expected to be present, and account for a compromise between a pre-specified clinically acceptable value of reliability and measurement error. Finally, although well-developed for continuous measurements, statistical methods for estimating the reliability and error of count measurements, as well as performing a meta-analysis of parameter estimates of reliability and measurement error reported across different primary studies, are currently lacking, and should be developed.

3. Evaluation of the reproducibility of grip strength measurements produced by hand-held digital dynamometers.

3.1. Introduction

Statistical methods that have been proposed for examining the reliability and measurement error of biomarkers were introduced in Chapter 2. In this chapter, these methods were applied in a cohort of 84 patients with sarcopenia and chronic inflammatory disease, including chronic liver disease (CLD), inflammatory Bowel Disease (IBD), inflammatory Rheumatoid Arthritis (IRA). The biomarker of interest is grip strength, which has been proposed for the evaluation of numerous medical conditions.

3.2. Clinical background

Evidence suggests that grip strength can function as a biomarker for the diagnosis, prognosis, and monitoring of numerous diseases that affect individuals to perform their daily activities and function independently [109-111]. The European Working Group on Sarcopenia in Older People (EWGSOP) recommends the use of grip strength measurement if sarcopenia (defined as progressive loss of muscle mass and function, most commonly occurring with ageing [112]) is suspected, with muscle quality subsequently confirmed by further investigations such as DXA, CT or MRI [111]. Furthermore, grip strength has been used for the evaluation of burn-affected upper limb strength [113], and for chronic conditions such as heart disease [114], diabetes [115, 116], arthritis [117], stroke [118], prostate cancer [119], and chronic obstructive pulmonary disease [120]. Yorke et al also showed a significantly negative association between grip strength and multi-morbidity [121], defined as the presence of two or more chronic conditions. Grip strength is most commonly assessed using a hand-

held dynamometer. The assessment is easy to perform, and recommendations on the testing procedures and the body positioning of the participants being tested have been available since 1981 [122]. During the assessment, the participants hold the dynamometer in the tested hand and squeeze with maximum isometric effort [122, 123]. The produced measurements most commonly express the amount of static force (kilograms, pounds or Newtons), or can also express the force per palmar surface area (millilitres of mercury or pounds per square inch), depending on the type of dynamometer being used [124], with higher measurements corresponding to a better health outcome. In order to be useful in clinical and research settings, the measurements of grip strength must be reliable and produced with low measurement error.

3.3. Study design

The data was collected as part of an observational, repeated measures study, which was conducted within the University Hospitals of Birmingham Foundation Trust. Each patient attended two clinic visits, with a different clinician assessing the patient at each visit. The two clinic visits were scheduled two weeks apart. This time interval was adopted by the study investigators as unlikely to observe any changes in the health status of the patients. Within each visit, three consecutive measurements of grip strength per hand (six in total) were obtained from each patient. The measurements were produced using a Takei digital dynamometer, with each measurement followed by a rest period of at least 30 seconds. Prior to the three analysis measurements recorded at each visit, clinicians provided verbal instructions to the patients. Vocal encouragement was also given to the patients during each attempt.

3.4. Components of variability in the measurements

Figure 3.1 illustrates the flow of the patients within the study. As depicted in Figure 3.1, there are potentially three components of variability in the measurements of grip strength, which were identified in discussion with the clinicians:

- **Between-patient variability.** That is, the differences in grip strength between the recruited patients. Ideally, this should be high compared to any other potential component of variability, as this indicates the ability of biomarkers to differentiate patients with a better grip strength to those with a worse.
- **Between-visit variability.** This refers to true differences in the way each patient performed at each visit, and can be understood as the first component of error in the measurements. Potential sources of between-visit variability include:
  - **1)** Random variability occurring within each visit (e.g., how the patients are feeling on testing day affecting their performance).
  - **2)** Systematic variability occurring between the two visits. For example, patients may perform better at the second visit due to their experience in the first (i.e., learning effect being present), or worse, as they may have experienced a change in health status between the first and second visit (although hypothesised against).
  - **3)** Interaction between the patient and the clinician assessing the patient at each visit (e.g., the instructions given by one clinician may motivate a patient more than the instructions given by the other clinician).

  However, variability due to **(2)** and **(3)** were expected to be low due to the simplicity of the procedure, the clear instructions provided to the patients prior to the testing sessions, and the change in the health status of the patients within two weeks being considered unlikely. Therefore, **(1)** was expected to be the main source of between-visit variability.

- **Between-measurement within-visit variability.** This can be thought as the second component of error in the measurements, and may be potentially attributed to:
  1) The random error occurring within the measurements due to inherent variation that the Takei dynamometer may have.
  2) Systematic differences in the performance of the patients. For example, the first measurement may be higher than the subsequent measurements as fatigue may set in after the first attempt, or lower, as patients may become more familiar with the procedure.

  However, **(1)** was considered to be the main source of between-measurement within-visit variability, as the clear instructions provided to the patients prior to each testing session, as well as the break between measurements that the patients were allowed, were expected to keep variability due to **(2)** at a low level.

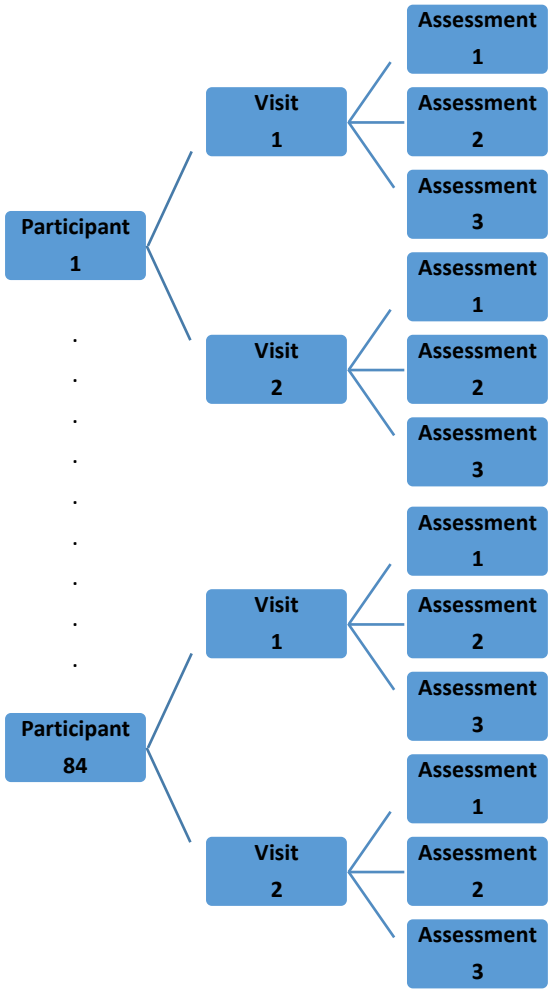
3.5. Aim and objectives

The aim of the chapter is to illustrate how standard statistical methods used for examining the reproducibility of biomarkers are applied, as well to provide evidence of the reproducibility of the Takei digital dynamometer, when used to evaluate grip strength. The two objectives of this chapter were:

(i) To examine the reproducibility of the grip strength measurements made at two different clinic visits, scheduled two weeks apart.

(ii) To examine the reproducibility of the grip strength measurements made within the same visit.

Figure 3.1. Study design.



3.6. Statistical methods

The baseline characteristics of the 84 patients recruited in the study were summarised using appropriate summary statistics. Two random effects linear models (one for each tested hand) were employed to estimate the three components of variability (see 3.4) in the measurements of grip strength. As described in Chapter 2, the use of this method requires the measurements produced at each potential level of variability to follow a normal distribution. In order to assess the distribution of the measurements produced at the patient level, the mean value of the multiple measurements produced at both visits was calculated for each patient, and used as the best estimate of the true grip strength value. Histograms were then used to assess whether the mean values produced from

each hand were normally distributed. If not, the measurements produced from each hand were log-transformed prior to the analysis. The assessment of normality for the remaining levels of variability (i.e., between and within-visit levels) was not possible due to the limited numbers of visits and measurements performed within the visits (two and three respectively).

3.6.1. Description of statistical model used to estimate the components of variability

Two linear regression models (one for each tested hand) were used to estimate the components of variability in the grip strength measurements. Each model was expressed as

$$y_{ijk} = \mu + u_k + v_{jk} + e_{ijk} \qquad (3.1)$$

where y_{ijk} denotes the i_{th} measurement of grip strength $(i = 1,2,3)$ produced within the j_{th} visit $(j = 1,2)$ from the k_{th} patient $(k = 1, \dots ,84)$, μ is the regression intercept, $u_k \sim N(0,\ \sigma_P)$ is the patient-level random effects parameter for the variability across the patients, $v_{jk} \sim N(0,\ \sigma_V)$ is the visit-level random effects parameter for the variability between the two visits, and $e_{ijk} \sim N(0,\ \sigma_I)$ is the random error term. The four parameter estimates obtained from the above regression model were:

- The grand mean of the produced measurements of grip strength, which equals the regression intercept (μ).
- The standard deviation of the patient-level random effects (σ_P).
- The standard deviation of the visit-level random effects (σ_V).
- The standard deviation of the individual measurements of grip strength produced from each patient within each visit (σ_I).

The models were fitted using the `xtmixed' command in Stata version 17, with restricted estimation of maximum likelihood (REML) [48].

3.6.2. Regression-based parameters of reliability and measurement error

Several statistical parameters were then calculated based on the above model-based estimates, in order to examine the reliability and error of the grip strength measurements produced from each hand. The intra class correlation was used to evaluate the reliability of the measurements (i.e., the ability of the test to distinguish patients despite any measurement error being present), while parameters of measurement error included (again) the intra class correlation, the standard error of measurement, the smallest detectable change, the coefficient of variation, and the reference change values. The calculation of each parameter is described in sections 3.6.2.1-3.6.2.5. All parameter estimates were presented along with 95% confidence intervals, which were constructed via multilevel bootstrapping with bias-correction (method described in Chapter 2) [79, 80, 125, 126]. For each parameter, a sampling distribution was obtained by fitting the model to 1000 bootstrapped samples [79, 126]. The 2.5% and 97.5% percentiles of the produced sampling distribution were used as the lower and upper confidence bound, respectively, and were adjusted in the appropriate direction in case the original model-based estimate did not lie at the 50th percentile.

3.6.2.1. Calculation of the intra class correlation

When used for evaluating the reliability of biomarkers, the intra class correlation (ICC) represents the proportion of the variability in the measurements that is attributed to true differences between the patients [15, 127]. In this example, this is mathematically expressed as

$$ICC_P = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_V^2 + \sigma_I^2}, \tag{3.2}$$

where σ_P^2 is the between-patient variance, σ_V^2 is the between-visit variance, and σ_I^2 is the between-measurement within-visit variance, with $(\sigma_V^2 + \sigma_I^2)$ representing the total error variance.

Furthermore, two additional intra class correlation coefficients were calculated in order to estimate the proportion attributed to between-visit (ICC_V) and within-visit variability (ICC_I), respectively, similarly by dividing the corresponding variance estimates by the total variance ($\sigma_P^2 + \sigma_V^2 + \sigma_I^2$). If skewness in the grip strength measurements was observed, the three ICC's were calculated in a similar fashion using the log-transformed measurements.

3.6.2.2. Calculation of the standard error of measurement

The standard error of measurement (SEM) equals the standard deviation of the repeated measurements made within the same patient, which reflects the spread of the repeated measurements around the patient's true score [15, 16, 127]. As described under section 3.4, there are two components of within-patient variability (i.e., error) in the measurements: (i) that due to differences between measurements made within the same visit, and (ii) that due to differences between measurements made at different visits.

For measurements produced from the same patient within the same visit, the standard error of measurement was equal to the corresponding estimate of the standard deviation ($SEM_I = \sigma_I$). This reflects the error that is expected for measurements produced from the same patient within each visit, due to both (i) the random error occurring within measurements, and (ii) any systematic differences in the performance of the patients within the visits. For measurements produced from the same patient at different visits, the calculation of the standard error of measurement was extended to

$$SEM_V = \sqrt{\sigma_V^2 + \sigma_I^2} \tag{3.3}$$

This reflects the total error that is expected for measurements produced from the same patient at different visits, due to both true between-visit differences in the performance of the patient, and

any additional variability arising from (i) and (ii). However, the calculation of SEM_V was omitted in case the distribution of the produced measurements was skewed, as this parameter is expressed in the original scale (kgs) and there is no more a natural interpretation of the produced estimate when the log-transformation is applied.

3.6.2.3. Calculation of the smallest detectable change

The smallest detectable change reflects the quantity below which the absolute difference between two measurements produced from the same patient is expected to lie with 95% probability, due to measurement error rather than a true change in the health status of the patient [15, 51]. For measurements produced from the same patient at different visits, the smallest detectable change was calculated as

$$SDC_V = \sqrt{2} \times 1.96 \times SEM_V \tag{3.4}$$

In order to identify any patients that were likely to have experienced a change in the health status in between the visits, the mean of the measurements produced from each patient was calculated for each visit. The absolute difference between the two mean values produced for each visit was in turn calculated and compared to the smallest detectable change, with values exceeding the smallest detectable change being considered as indicative of a change in the health status. For measurements produced from the same patient within the same visit, the smallest detectable change was not calculated. This was because true changes in the health status of the patients were not expected to occur in between measurements taken within the same visit, as the measurements were taken successively.

3.6.2.4. Calculation of the coefficient of variation

The coefficient of variation reflects the (percentage) spread of repeated measurements made within the same patient, around the grand mean [15, 16]. It is calculated as the ratio of the standard deviation of the repeated measurements made within patients, to the grand mean of the measurements. For measurements produced from the same patient within the visits, the coefficient of variation was equal to

$$CV_I = \frac{SEM_I}{\mu} = \frac{\sigma_I}{\mu} \tag{3.5}$$

For measurements produced from the same patient at different visits, the calculation of the coefficient of variation was extended to

$$CV_V = \frac{SEM_V}{\mu} = \frac{\sqrt{\sigma_V^2 + \sigma_I^2}}{\mu} \tag{3.6}$$

This reflects the spread of measurements produced from the same patient at different visits, around the grand mean, due to true between-visit differences in the performance of the patient, as well as any additional within-visit variability. If skewness in the grip strength measurements was observed, the coefficient of variation was calculated using the formula provided by Cole [61], as

$$CV = \sqrt{e^{(SEM^{log})^2} - 1}\,, \tag{3.7}$$

where SEM^{log} is the standard error of measurement estimated on the logarithmic scale.

3.6.2.5. Calculation of the reference change value

The reference change value has been proposed for laboratory-based biomarkers with low individuality (see Chapter 2), and aims to define the minimal percentage change from the first measurement that needs to be exceeded, in order to state that a true change in the health status of the individual has taken place. Although this study is not laboratory-based, the chapter aims to

illustrate how the parameter is applied in clinical practice. Like the smallest detectable change, the parameter was only used for measurements taken at different visits, as true changes in the health status of the patients were not expected to occur in between successive measurements taken within the same visit. The reference change value was calculated as

$$RCV_V = \sqrt{2} \times 1.96 \times CV_V, \tag{3.8}$$

where CV_V is calculated as described in section 3.6.2.4. In order to define the absolute difference between two within-individual measurements that needs to be exceeded for a significant change, the value produced from equation 3.8 was in turn multiplied by the mean of the grip strength measurements produced at the first visit.

3.6.3. Additional parameters used for pairwise comparisons

Additional analyses involved the calculation of the limits of agreement as a parameter of measurement error, and the calculation of the Pearson correlation as a parameter of reliability. As described in Chapter 2, these parameters can only be used when two measurements are available for each patient. For examining the variability between the two visits, both the highest and average of the three measurements produced from each patient within each visit were used as summary measures in the analysis. This gives a total of 2 (summary measures) ×2 (hands) = 4 pairwise comparisons between the two visits. For examining within-visit variability, each pairwise comparison available within each visit (i.e., first v second measurement, first v third measurement, second v third measurement produced from each patient) was performed. This gives a total of 3 (pairs within visit) ×2 (visits) ×2 (hands) = 12 comparisons. Although the comparisons were likely to be correlated, no adjustment for multiple comparisons was made, as no hypothesis testing was performed.

3.6.3.1. Calculation of the limits of agreement

The limits of agreement were calculated as the interval of twice the standard deviation of the differences between two repeated measurements, either side of the mean difference [15, 47]. In mathematical notation this is expressed as

$$LoA = d \pm 1.96 \times SD_d\,, \tag{3.9}$$

where d represents the systematic differences between two repeated measurements, and SD_d represents the random variability occurring within the two measurements. The produced interval was presented along with the lower 95% confidence bound of the lower limit of agreement, and the upper 95% confidence bound of the upper limit of agreement, as recommended by Bland and Altman [47]. These were calculated as

$$95\%CI - LoA_{U/L} = (d \pm 1.96 \times SD_d) \pm t(n_{pairs} - 1, 0.025)\sqrt{\frac{3 \times {SD_d}^2}{n_{pairs}}}, \tag{3.10}$$

where n_{pairs} is the number of available pairs of measurements, and $t(n_{pairs} - 1, 0.025)$ is the critical value of the Student's t distribution with $n_{pairs} - 1$ degrees of freedom and 2.5% significance level.

As described in Chapter 2, this method requires the following two assumptions to be met: (i) the differences between the two measurements made within-patients should be normally distributed, and (ii) there should be no evidence of heteroscedasticity (i.e., the within-patient variability should not be increasing with higher values of grip strength). Therefore, for all different comparisons performed, the first assumption was assessed using histograms, while the latter was assessed with Bland-Altman plots, plotting the difference between each pair of measurements available for each patient against the corresponding mean of the measurements. If either assumption was violated, the measurements were log-transformed prior to the analysis, as recommended by Bland and Altman [47]. In this case, the limits of agreement were expressed as a function of the mean of a randomly selected pair of measurements produced from the same patient $(\bar{Y})$, as described in Euser et al [56].

For a given value of $\bar{Y} = \frac{Y_1+Y_2}{2}$, the difference between two randomly selected measurements made on the same patient $(Y_1 - Y_2)$ lies within

$$LoA = \pm 2\bar{Y}(10^a - 1)/(10^a + 1), \quad (3.11)$$

where $a = 1.96 \times SD_d$.

3.6.3.2. Calculation of the Pearson correlation coefficient

The Pearson correlation coefficient was calculated as

$$r = \frac{\sum_{i=1}^{n}(y_{i1}-\overline{y_1})(y_{i2}-\overline{y_2})}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\overline{y_1})^2 \sum_{i=1}^{n}(y_{i2}-\overline{y_2})^2}}, \quad (3.12)$$

where y_{i1} and y_{i2} are the measurements of the i_{th} patient $(i = 1, \dots, n)$ at the first and second testing occasion, respectively, and $\overline{y_1}$ and $\overline{y_2}$ are the corresponding mean values across all patients. A 95% confidence interval for r was based on the Fisher's Z transformation [67], and constructed via

$$(r_L, r_U) = \left(\frac{e^{2Z_L}-1}{e^{2Z_L}+1}, \frac{e^{2Z_U}-1}{e^{2Z_U}+1}\right), \quad (3.13)$$

where $Z_{U/L} = Z \pm 1.96\sqrt{\frac{1}{(n-2)}}$ is the upper/lower 95% confidence bound for the Fisher's Z statistic, with $Z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$.

As described in Chapter 2, the use of the Pearson correlation coefficient as a reliability parameter requires the following assumptions to be met: (i) the measurements produced at the first and second testing occasions should both be normally distributed, (ii) the relationship between the measurements produced at each testing occasion should be linear (iii) no extreme outliers are present, (iv) no systematic differences between the two testing occasions are observed. For all different comparisons performed, (i) was assessed using histograms, while (ii)-(iv) were examined through scatterplots, plotting the measurements produced at the first testing occasion against those

produced at the second testing occasion (e.g., mean of first visit v mean of second visit). If either assumption was violated, the correlation was calculated using the Spearman's non-parametric test, as

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n} (rank[y_{i1}] - rank[y_{i2}])^2}{n(n^2 - 1)}, \tag{3.14}$$

where $rank[y_{i1}]$ and $rank[y_{i2}]$ are the ranks of y_{i1} and y_{i2}, respectively.

3.7. Results

The baseline characteristics of the 84 patients recruited in the study are presented in Table 3.1. The patients were on average 55.0 (SD=10.2) years old, with a mean BMI of 29.6 (SD=6.5). The number of right-handed and female patients was 71 (85.0%) and 29 (34.0%), respectively. The majority (n=53, 63.1%) were patients with chronic alcohol-related liver disease. Of the remaining 31, 11 (13.1%) had inflammatory Bowel Disease, another 11 (13.1%) had inflammatory Rheumatoid Arthritis, and 9 (10.7%) had non-alcoholic related liver disease. All patients (100%) attended the first visit, with 70/84 (83%) patients additionally attending the second visit. Of the 70 patients attending both visits, one performed 2 out of 3 measurements per hand during the second visit, two did not perform any left hand-based measurements during the first visit, while an additional patient did not perform any left hand-based measurements at either visit (see Figure 3.2). The frequency distributions of the mean of the multiple measurements produced from each hand are displayed in Figure 3.3. The distribution of the mean values produced from the right and left hand was approximately normal, with a mean value (SD, n) of 29.57 (10.30, 84) and 28.60 (9.89, 83), respectively.

Figure 3.2. Flow chart of patients, showing the numbers who provided data at each visit.

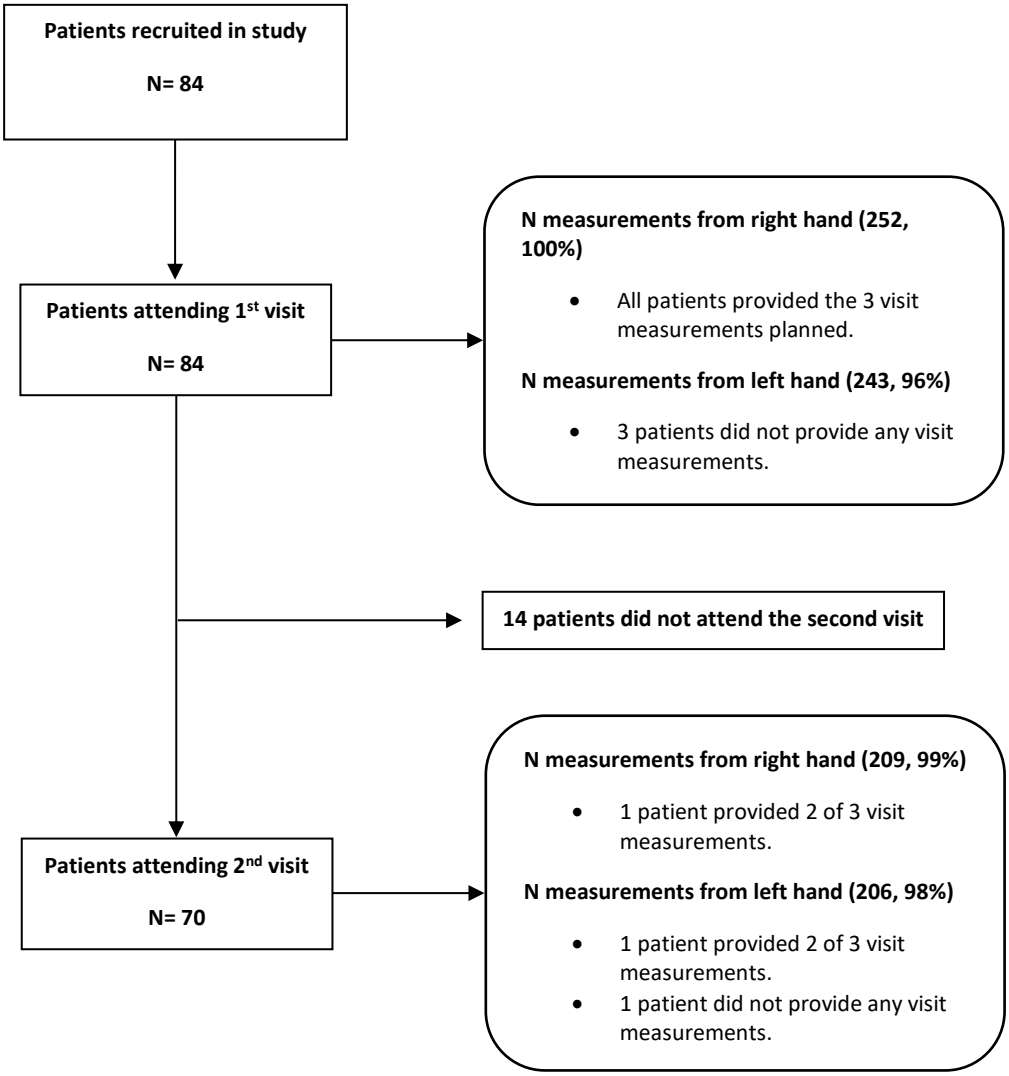
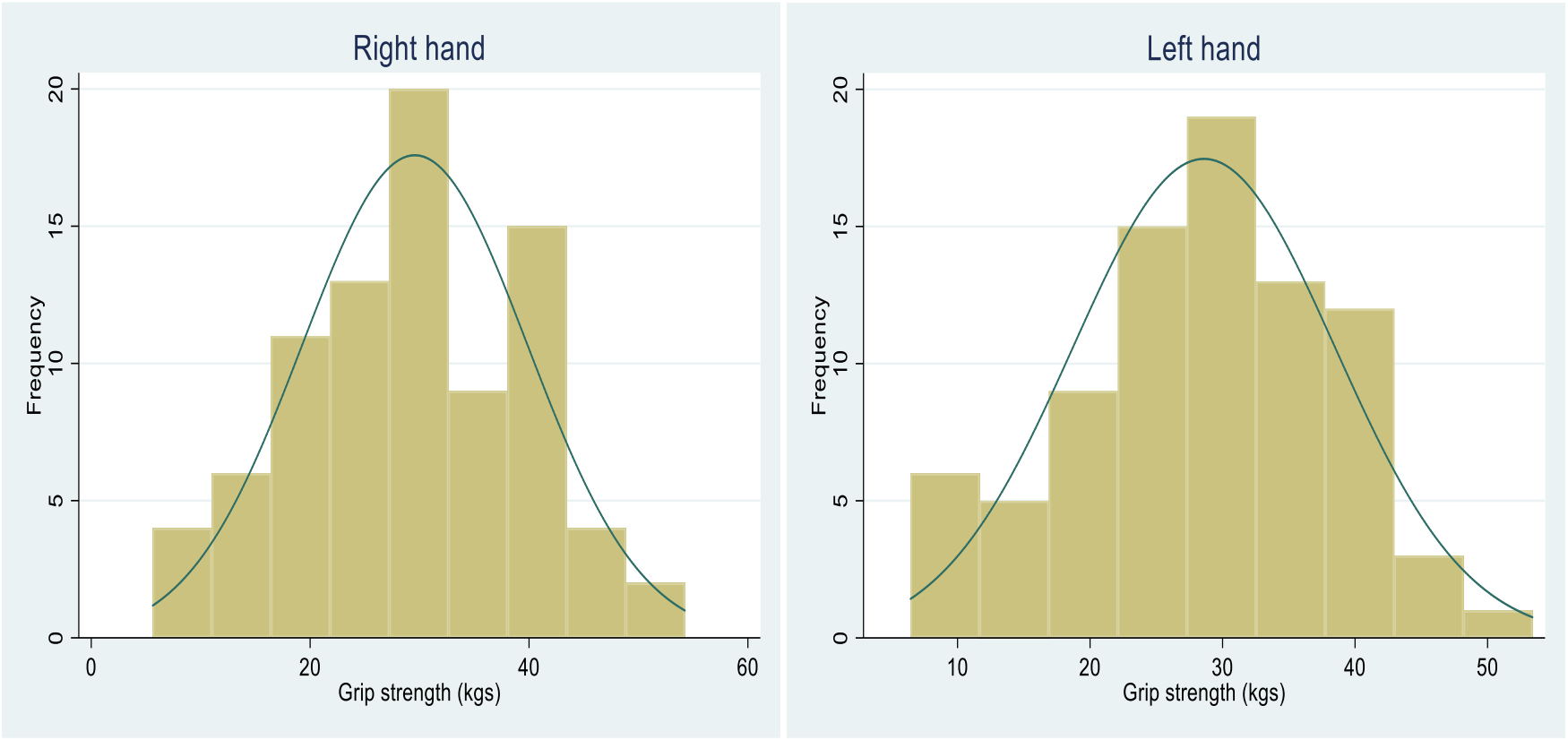


Table 3.1. Baseline characteristics of 84 patients recruited in the study.

Age, mean (SD)	55.0 (10.2)
Female patients, N (%)	29 (34.0%)
Medical condition, N (%)	
Chronic alcohol-related liver disease	53 (63.1)
Inflammatory Bowel Disease	11 (13.1)
Inflammatory Rheumatoid Arthritis	11 (13.1)
Chronic non-alcohol related liver disease	9 (10.7)
Right-handed patients, N (%)	71 (85.0%)
BMI, mean (SD)	29.56 (6.49)

Figure 3.3. Distribution of the mean of multiple measurements produced from each patient, by hand.



3.7.1. Regression-based parameters of reliability and measurement error

The results obtained from each hand are presented in Table 3.2. This section only describes the results obtained from the right hand, as both hands yielded very similar results.

3.7.1.1. Estimates obtained from the linear random effects model

The estimate for the regression intercept, which also corresponds to the mean value of the grip strength measurements, was 29.58kgs. The estimate for the between-patient standard deviation was as large as 10.16kgs, indicating high variability in the grip strength measurements from one patient to another. The estimates of the between-visit and between-measurement within-visit standard deviation was substantially lower compared to the between-patient standard deviation estimate, and very similar to each other (1.85kgs and 2.13kgs respectively).

3.7.1.2. Results produced for the intra class correlation coefficient

The patient-level estimate of the intra class correlation was equal to 0.928. This denotes that 92.8% [95% CI: (90.1%, 95.0%)] of the total variability in the measurements of grip strength is attributed to true differences between patients. Of the remaining 7.2%, 3.1% [95% CI: (1.0%, 5.0%)] was attributed to true differences in the way a patient performed at the two different visits, and 4.1% [95% CI: (3.7%, 5.4%)] was attributed to differences between measurements produced from the same patient within the same visit.

3.7.1.3. Results produced for the standard error of measurement

For measurements produced from the same patient within the same visit, the standard error of measurement was equal to 2.13kgs. This reflects a 68% chance of a measurement repeated within the same visit lying within 2.13kgs [95% CI: (2.08, 2.28)] of the true score of the patient, due to (i)

the patient performing differently between measurements made within the same visit, and (ii) due to the random error occurring within each measurement. For the measurements produced from the same patient at two different visits, the standard error of measurement was equal to 2.82kgs. This reflects a 68% chance of a measurement repeated at a different visit lying within 2.82kgs [95% CI: (2.36, 3.30)] of the true score of the patient due to true between-visit differences in the performance of the patient, as well as any additional variability attributed to (i) and (ii).

3.7.1.4. Results produced for the smallest detectable change

The produced estimate for the smallest detectable change was equal to 7.81 [95% CI: (6.54, 9.14)]. If the difference between two measurements produced from the same patient at two different visits exceeds 7.81kg, it can be stated with 95% confidence that this difference reflects a true change in the health status of the patient, rather than one anticipated due to measurement error. A difference larger than 7.81kg was observed in two patients (absolute difference between visit means was 8.33kgs and 9.40kgs), which indicates that these patients were likely to have experienced a change in their health status.

A difference larger than 7.81kg was observed in two patients (absolute difference between means produced at each visit was 8.33kgs and 9.40kgs), which indicates that these patients were likely to have experienced a true change in grip strength.

3.7.1.5. Results produced for the coefficient of variation

For measurements produced from the same patient within the same visit, the coefficient of variation was equal to 0.072. This reflects a 68% chance of a measurement repeated within the same visit lying within 7.2% [95% CI: (7.1%, 7.8%)] of the grand mean, due to (i) the patient performing differently between measurements made within the same visit, and (ii) due to the random error

occurring within each measurement. For measurements produced from the same patient at two different, the coefficient of variation was equal to 0.095. This reflects a 68% chance of a repeated measurement lying within 9.5% [95% CI: (8.0%, 11.2%)] of the grand mean due to true between-visit differences in the performance of the patient, as well as any additional variability attributed to (i) and (ii).

3.7.1.6. Results produced for the reference change value

The low values obtained for the index of individuality (<0.6 for both hands) makes the use of the reference change value appropriate for defining a significant change occurring within the individuals. For the right hand, the reference change value was equal to 26.3%, while the mean value of the grip strength measurements produced at the first visit was 29.3kgs. One can then state that a difference larger than or equal to 29.3 × (26.3/100) = 7.71kgs reflects a significant change in the health status of the patients.

Table 3.2. Results obtained from linear regression analyses performed for each hand.

	Right hand	**Left hand**
- ***Model-based parameter estimates***		
Grand mean of measurements, μ (95% CI)	29.58 (28.43, 30.47)	28.61 (27.35, 29.42)
Between-patient standard deviation, σ_P (95% CI)	10.16 (9.39, 10.76)	9.76 (8.92, 10.37)
Between-visit standard deviation, σ_V (95% CI)	1.85 (1.16, 2.37)	1.75 (1.10, 2.23)

Between-measurement within-visit standard deviation, σ_I (95% CI)	2.13 (2.08, 2.28)	1.85 (1.80, 2.01)
- ***Parameter used to examine the reliability of the measurements***		
Intra class correlation (95% CI)	92.8% (90.1%, 95.0%)	93.6% (90.7%, 95.4%)
- ***Parameters of between-visit variability in measurements produced from the same patient***		
Standard error of measurement (95% CI)	2.82 (2.36, 3.30)	2.55 (2.17, 3.05)
Smallest detectable change[1] (95% CI)	7.81 (6.54, 9.14)	7.06 (6.01, 8.45)
Coefficient of variation (95% CI)	9.5% (8.0%, 11.2%)	8.9% (7.5%, 10.5%)
Reference change values (95% CI)	26.3% (22.2%, 31.0%)	24.7% (20.8%, 29.1%)
Index of individuality (95% CI)	0.28 (0.23, 0.33)	0.26 (0.22, 0.32)
- ***Parameters of within-visit variability in measurements produced from the same patient***		
Standard error of measurement (95% CI)	2.13 (2.08, 2.28)	1.85 (1.80, 2.01)
Coefficient of variation (95% CI)	7.2% (7.1%, 7.8%)	6.5% (6.3%, 7.1%)

[1]*Three patients presented a difference larger than the smallest detectable change (one for the right hand, one for the left hand, and one for both hands).*

3.7.2. Additional parameters used for pairwise comparisons

The results obtained from each pairwise comparison performed between and within the visits were presented in Tables 3.3 and 3.4, respectively. The assumptions required for the use of the limits of agreement and the Pearson correlation coefficient were met across all comparisons performed, particularly when the mean of multiple measurements was used as the visit summary measure (see Appendices A1-A4). The Bland Altman plots produced from the pairwise comparisons performed between visits are presented in Figure 3.4. Similar to section 3.7.1, only the results obtained from the right hand are described, as the results produced from the two hands were very similar to each other.

3.7.2.1. Variability between visits

Limits of agreement

When the mean of the three within-patient measurements produced at each visit was used as a summary measure, the limits of agreement produced for the right hand ranged from -6.86kgs to 5.53kgs. When the summary measure was altered to the highest of the three within-patient measurements produced at each visit, the limits of agreement ranged from -6.94kgs to 6.15kgs. For both summary measures, the observed variability between the paired values was mainly attributed to the random error occurring within the measurements, as the systematic differences between measurements were negligible (-0.66kgs for the mean, -0.39kgs for the highest of three measurements).

Pearson correlation coefficient

When the mean of the three within-patient measurements produced at each visit was used as a summary measure, the estimate for the Pearson correlation coefficient was as high as 0.96. A similar estimate (=0.95) was produced when the summary measure was altered to the highest of the three measurements produced from each patient at each visit.

3.7.2.2. Variability in the measurements produced within visits

Limits of agreement

The produced limits of agreement were similar across all pairwise comparisons performed within each visit, with narrower intervals observed for comparisons including adjacent measurements. The lower limit ranged from -8.13kgs to -5.09kgs, while the upper limit ranged from 4.06kgs to 7.14kgs. In all different comparisons performed, the observed variability between the paired measurements

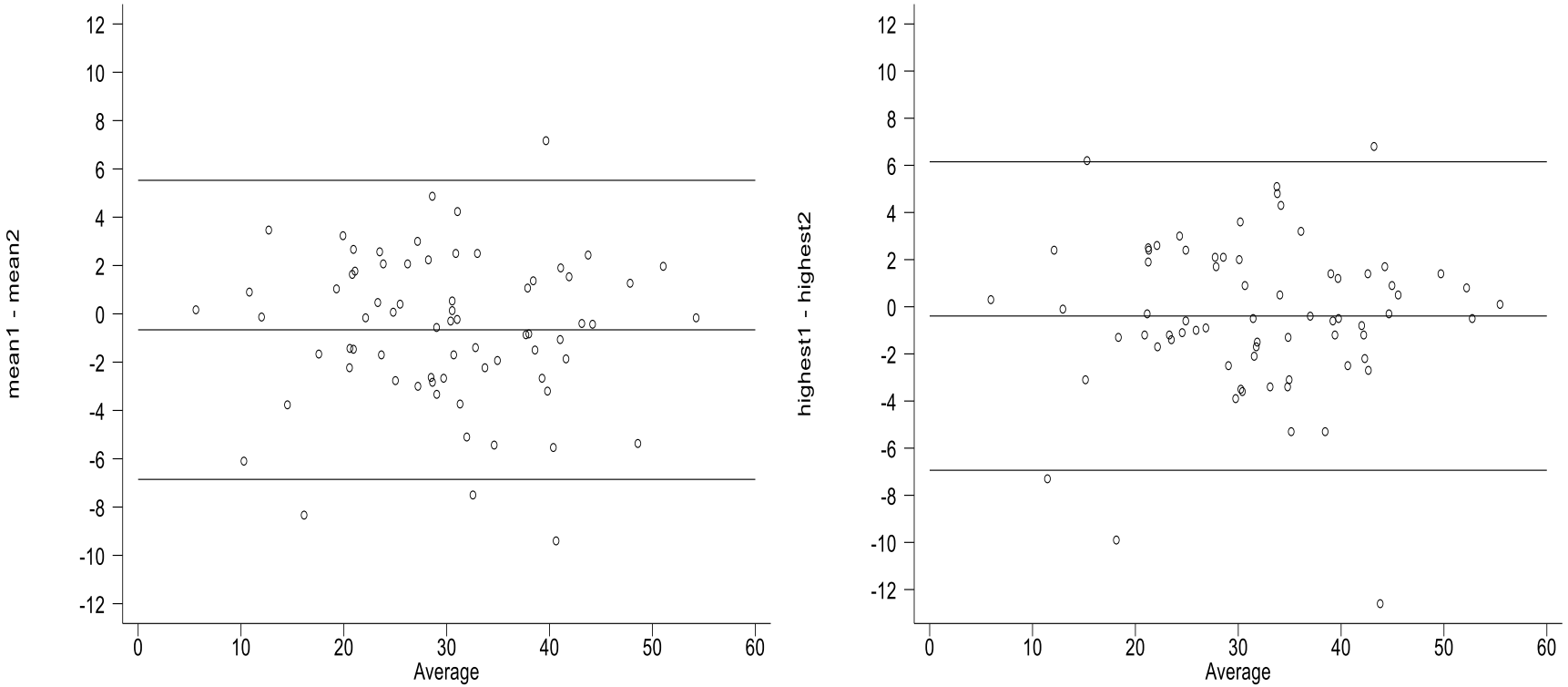
was mainly attributed to the random error occurring within the measurements, as the systematic differences between measurements were negligible (varying from -0.71kgs to 0.27kgs).

Pearson correlation coefficient

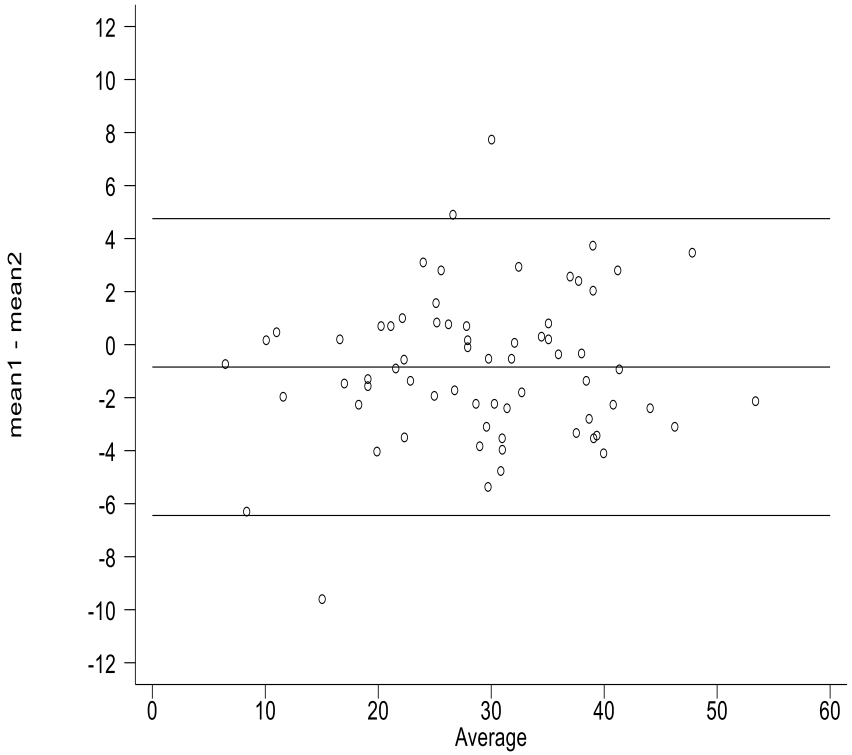
A high reliability was noted for the three pairwise comparisons made within each of the two visits, the estimates for the Pearson correlation coefficients ranging between 0.94 and 0.97.

Figure 3.4. Bland Altman plots of the mean and highest of three within-individual measurements produced at each visit.

i) Values produced from the right hand



ii) Values produced from left hand



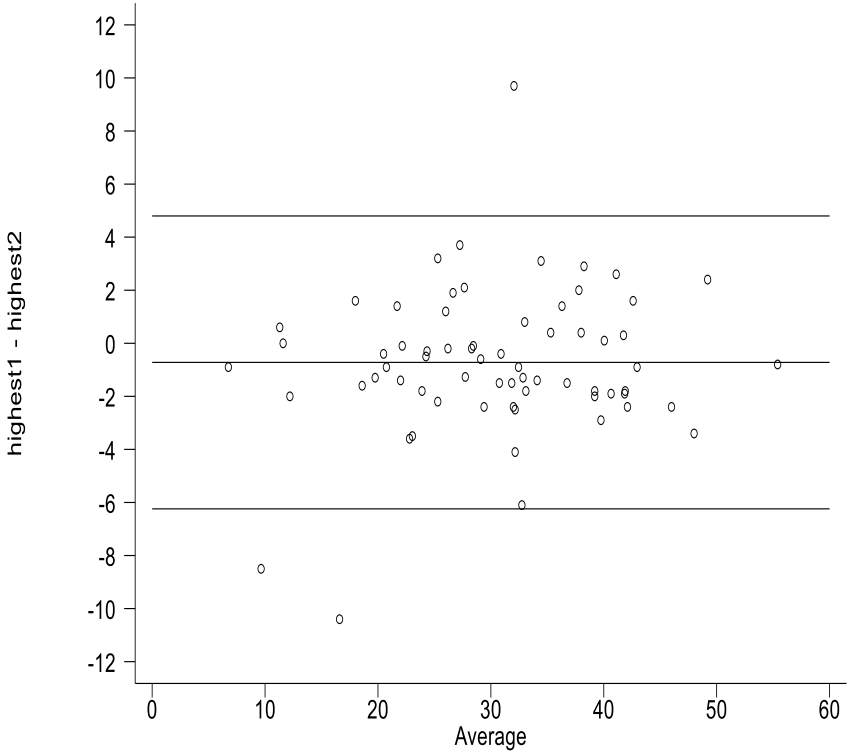


Table 3.3. Results obtained from pairwise comparisons performed between the two visits.

	Right hand	**Left hand**
- ***Mean of three measurements produced at each visit***		
Mean difference [(95% CI), N]	-0.66 [(-1.40, 0.07), 70]	-0.85 [(-1.53, -0.16), 67]
95% limits of agreement (95% CI)	-6.86 to 5.53 (-8.13, 6.81)	-6.44 to 4.75 (-7.63, 5.94)
Pearson correlation (95% CI)	0.96 (0.93, 0.97)	0.96 (0.94, 0.98)
- ***Highest of three measurements produced at each visit***		
Mean difference (95% CI)	-0.39 [(-1.17, 0.39), 70]	-0.72 [(-1.40, -0.05), 67]
95% limits of agreement (95% CI)	-6.94 to 6.15 (-8.29, 7.50)	-6.24 to 4.80 (-7.41, 5.96)
Pearson correlation (95% CI)	0.95 (0.93, 0.97)	0.96 (0.94, 0.98)

Table 3.4. Results obtained from pairwise comparisons performed within each visit.

	Right hand	**Left hand**
- ***Visit 1: measurement 1 v measurement 2***		
Mean difference [(95% CI), N]	-0.21 [(-0.84, 0.42), 84]	-0.86 [(-1.45, -0.28), 80]
95% limits of agreement (95% CI)	-6.00 to 5.59 (-7.09, 6.67)	-6.10 to 4.38 (-7.11, 5.39)
Pearson correlation (95% CI)	0.96 (0.94, 0.98)	0.97 (0.95, 0.98)
- ***Visit 1: measurement 1 v measurement 3***		
Mean difference [(95% CI), N]	-0.49 [(-1.32, 0.34), 84]	-0.76 [(-1.50, -0.02), 80]
95% limits of agreement (95% CI)	-8.13 to 7.14 (-9.56, 8.57)	-7.41 to 5.89 (-8.69, 7.17)
Pearson correlation (95% CI)	0.94 (0.90, 0.96)	0.95 (0.92, 0.97)
- ***Visit 1: measurement 2 v measurement 3***		
Mean difference [(95% CI), N]	-0.28 [(-0.81, 0.24), 84]	0.11 [(-0.41, 0.62), 81]

95% limits of agreement (95% CI)	-5.09 to 4.52 (-5.99, 5.43)	-4.53 to 4.74 (-5.41, 5.62)
Pearson correlation (95% CI)	0.97 (0.96, 0.98)	0.975 (0.96, 0.98)
- ***Visit 2: measurement 1 v measurement 2***		
Mean difference [(95% CI), N]	0.27 [(-0.50, 1.04), 70]	-0.22 [(-0.76, 0.32), 69]
95% limits of agreement (95% CI)	-6.17 to 6.71 (-7.50, 8.04)	-4.72 to 4.28 (-5.65, 5.21)
Pearson correlation (95% CI)	0.96 (0.93, 0.97)	0.974 (0.96, 0.98)
- ***Visit 2: measurement 1 v measurement 3***		
Mean difference [(95% CI), N]	-0.46 [(-1.18, 0.26), 69]	-0.33 [(-0.92, 0.27), 68]
95% limits of agreement (95% CI)	-6.44 to 5.53 (-7.68, 6.77)	-5.21 to 4.56 (-6.24, 5.59)
Pearson correlation (95% CI)	0.96 (0.94, 0.98)	0.97 (0.95, 0.98)
- ***Visit 2: measurement 2 v measurement 3***		
Mean difference [(95% CI), N]	-0.71 [(-1.28, -0.14), 69]	-0.12 [(-0.68, 0.43), 68]
95% limits of agreement (95% CI)	-5.48 to 4.06 (-6.47, 5.05)	-4.72 to 4.47 (-5.68, 5.43)
Pearson correlation (95% CI)	0.97 (0.96, 0.98)	0.97 (0.96, 0.98)

3.8. Discussion

This study aimed to examine the error of the grip strength measurements produced from patients from a Takei digital dynamometer, as well as the variability in the measurements produced from the same patient (i.e., measurement error), both between two different visits and within a single visit. For this purpose, 84 patients were recruited and assessed at two different clinic visits, set two weeks apart, with 3 measurements produced from each patient at each visit. However, this sample size was not based on any formal statistical calculation, while the limited number of visits and measurements performed within each visit did not allow the corresponding components to variability to be estimated with high precision. Furthermore,

this study was restricted to patients with sarcopenia and chronic inflammatory diseases, which does not allow generalisation of the results to a wider population.

The regression-based analysis indicated that the measurements produced from both hands were accompanied by error, which was attributed to both variability in the way patients performed between the two visits, and variability in the measurements produced for the same patient within each visit. However, for both hands, the measurements were produced with high reliability and acceptable levels of error. The produced estimates for the standard error of measurement were approximately equal to those reported in similar studies examining the reproducibility of grip strength measurements and were considered acceptable by the authors [33, 128], while the estimated coefficient of variation for both hands was less than 10%. The cut-off value of 10% was used as an indicator of acceptable within-patient variability in similar studies assessing the reproducibility of grip strength measurements [108, 129]. Furthermore, the proportion of the total variability attributed to either between-visit variability or between-measurement within-visit variability appeared negligible. In contrast, the intra class correlation for the patient level was >0.90 for both hands, indicating that more than 90% of the total variability in the grip strength measurements was attributed to true differences between patients. This reveals that the produced measurements were highly reliable in distinguishing patients with a stronger grip, from patients with a weaker grip, despite the presence of measurement error. Only 3 patients presented an absolute between-visit difference larger than the smallest detectable change, with 2 out of 3 patients showing this difference in either hand (right or left), and the remaining patient showing this difference in both hands. Based on this method, it can be stated with 95% confidence that these patients were likely to have experienced a change in their health status. As expected, the results obtained from the smallest detectable change were very similar to those obtained from the reference change value.

Additional analysis included the use of the limits of agreement as a parameter of measurement error, and the Pearson correlation coefficient as a parameter of reliability. An obvious disadvantage of the two parameters is that both can only be used when a pair of measurements is available from each patient. Furthermore, the Pearson correlation coefficient may be used as a reliability parameter only when the variability between the paired measurements is attributed to the random error occurring within each measurement, as it is unable to account for any systematic variability between the two measurements [15]. However, the Pearson correlation coefficient was in this case expected to be an adequate measure of reliability, given that: (i) the individual effect of clinicians assessing the patients was in this case expected to be negligible (measurements are produced from an objective device with no clinical interpretation required), and (ii) every possible attempt was made to minimise any other potential source of systematic variability (as described under section 3.4). The results confirmed that, for all pairwise comparisons performed between and within visits, the variability between paired measurements was mainly attributed to the random error occurring within measurements, as systematic differences between measurements were low (as expected with an objective test), while high reliability estimates were observed across all pairs (Pearson correlation values>0.90).

3.9. Conclusion

The results indicated that, although not error-free, measurements produced from a Takei digital dynamometer are highly reliable. However, in order to state that Takei digital dynamometers can be used in clinical practice as part of evaluating loss of grip strength in patients with sarcopenia and chronic inflammatory disease, new evidence is required based on a formal sample size calculation (e.g., number of patients required to detect a pre-specified clinically acceptable estimate of reliability and/or measurement error), and a higher number of within-individual measurements. However, the latter may not always be possible when a frail population is tested. Finally, the results are only applicable to patients with sarcopenia and chronic conditions, and should not be generalised to a wider population.

4. Current state of systematic review methods and meta-analytic approaches used for evaluating the reliability and measurement error of biomarkers

4.1. Introduction

A systematic review involves searching the medical literature in order to collect, critically appraise, and synthesize results reported in primary studies. It uses explicit, systematic methods which are selected with a view to eliminating bias, thus providing more robust findings [130, 131]. The conclusions drawn can be used for research purposes, guideline development, evidence-based patient care and policy-making [15]. While systematic reviews examining the diagnostic accuracy or prognostic ability of biomarkers have become commonplace [87, 132], the extent to which systematic methods have been adopted for systematic reviews of the reliability and measurement error of biomarkers is less clear. A 2009 methodological review of the measurement properties of health status measurements found the 148 included reviews to be generally of poor quality [133]. For example, 22% of the identified reviews used only one electronic data base for identifying primary studies, while the search strategy was often too narrow or not clearly described. The majority of the reviews (i.e., >70%) did not report the approach to article selection and data extraction, while in some reviews, these tasks were completed by a single reviewer (rather than at least two independent reviewers). The subsequent COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) initiative has since provided a guide for conducting systematic reviews of patient-reported outcome measures [134]. Although these standards may not be fully applicable to biomarkers, guidelines provided on the domains of reliability and measurement error of patient-reported outcome measures are expected to

have also improved the methodological quality of reviews examining the reliability and measurement error of biomarkers.

Furthermore, systematic reviews typically use meta-analytical methods to combine the estimates of a particular parameter of interest reported in the identified primary studies, so that a summary estimate is obtained. Such meta-analytical methods are well established for estimates regarding the diagnostic accuracy or prognostic ability of biomarkers [87, 88], but less developed with respect to their reliability and measurement error. Thus, this chapter purposely identified published systematic reviews examining the reliability and measurement error of biomarkers, in order to appraise the review process used in the identified systematic reviews, and examine the current state of statistical methods used for the meta-analysis of any parameters of reliability and measurement error.

4.2. Objectives

The primary objective of this chapter was to examine the current state of statistical methods used for the meta-analysis of parameters of reliability and measurement error. A secondary objective was to evaluate the review process used in the identified systematic reviews. For the primary objective, the identified meta-analytic methods were summarised as frequency of use (and percentage), while the appropriateness of use was also discussed. For the secondary objective, each of the key steps of the review process were examined, including the comprehensiveness of the search strategy, the presence and nature of quality assessment, and use of independent screening, data extraction and quality assessment.

4.3. Methods

Data sources

Electronic searches of MEDLINE and EMBASE were undertaken by one reviewer (JD) on 4-4-2019 to identify relevant English language studies published to date. The search strategies were informed by the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) search filter for identification of studies on measurement properties [135], and are provided in Appendix B1. The starting date was set at 2010 to ensure the retrieval of a sufficient number of eligible reviews that reflect the current standard of systematic reviews in the field, and because the previous COSMIN review on a similar topic was published in 2010.

Study selection

Study selection (title and abstract and full text) was undertaken by one of the reviewers involved in this work (KT or JD) after an initial pilot on a sample of 60 reviews to ensure satisfactory agreement in study selection between the 2 reviewers. Systematic reviews were eligible for inclusion if:

- the reviews were examining any potential sources of variability of at least one test in any population.
- the reviews were published in English.
- at least one electronic database was used for the identification of primary studies.
- at least one statistical parameter of reliability and measurement error was reported.

Reviews only available as conference abstracts or protocols were excluded. No other restrictions were applied. All excluded reviews were double checked, and any disagreements

were resolved by consensus between the two reviewers. A relatively broad definition of a systematic review was used to provide a comprehensive reflection of current practice.

Data collection

Data extraction was carried out by one reviewer (KT or JD) using a pre-specified data extraction form (provided in Appendix B2), which was piloted on a sample of 10 randomly selected reviews. Detailed information was extracted regarding:

- the type of test and the condition it was tested for.
- The sources of test variability examined. All reported sources of variability were of interest and recorded (e.g., inter or intra-observer variability).
- the review methods used, including the literature search and approach to screening, data extraction, and quality assessment.
- the statistical methods used to examine the reproducibility of biomarkers.
- the statistical methods used for pooling the reported estimates of reproducibility.

Appraisal of review process

To appraise the quality of the review process, the following information was recorded:

- whether the inclusion and exclusion criteria for primary studies were described.
- whether the characteristics of the included primary studies were presented.
- whether the full search strategy was provided.
- whether at least one more database was searched in addition to PUBMED/ MEDLINE.
- whether an assessment of the methodological quality of the primary studies included in the reviews was conducted.

- whether article selection, data extraction, and quality assessment were carried by at least two independent reviewers.

The choice of the aforementioned criteria was based on a methodological review of the measurement properties of health status measurements, conducted by Mokkink et al in 2009 [133].

Data synthesis

The extracted information was summarised as frequencies and percentages. A narrative synthesis was then used to assess the quality of the review process, and the appropriateness of the statistical methods employed for the meta-analysis of the reported estimates.

4.4. Results

4.4.1. Summary of Reviews identified

A total of 3284 unique records were retrieved, of which 279 were selected for full text assessment. Of the 279 records, 219 met the eligibility criteria (Figure 4.1). A summary of the identified reviews is presented in Table 4.1, while the characteristics of each individual review are presented in Appendix B3. A wide range of target conditions was covered, from severe types of diseases including rheumatic disorders (15, 7%), heart disease (14, 6%), and various types of cancer (8, 4%), to the physical performance of healthy populations (15, 7%). Approximately half (114, 52%) of reviews focused on assessment of physical performance (using device-based and/or non-device based tests), 62 (28%) evaluated imaging tests, and the remainder evaluated laboratory biomarkers (13, 6%), physiologic measures (12, 5%), clinical examination (8, 4%), or multiple test types (11, 5%). Of the 219 reviews, the observer effect was examined in 138 reviews (63%). Of these 138 reviews, the majority (123, 89%) examined

both inter and intra-observer variability, with an additional 10 (8%) and 5 (4%) reviews only examining inter-observer or intra-observer variability, respectively.

Figure 4.1. PRISMA flow chart.

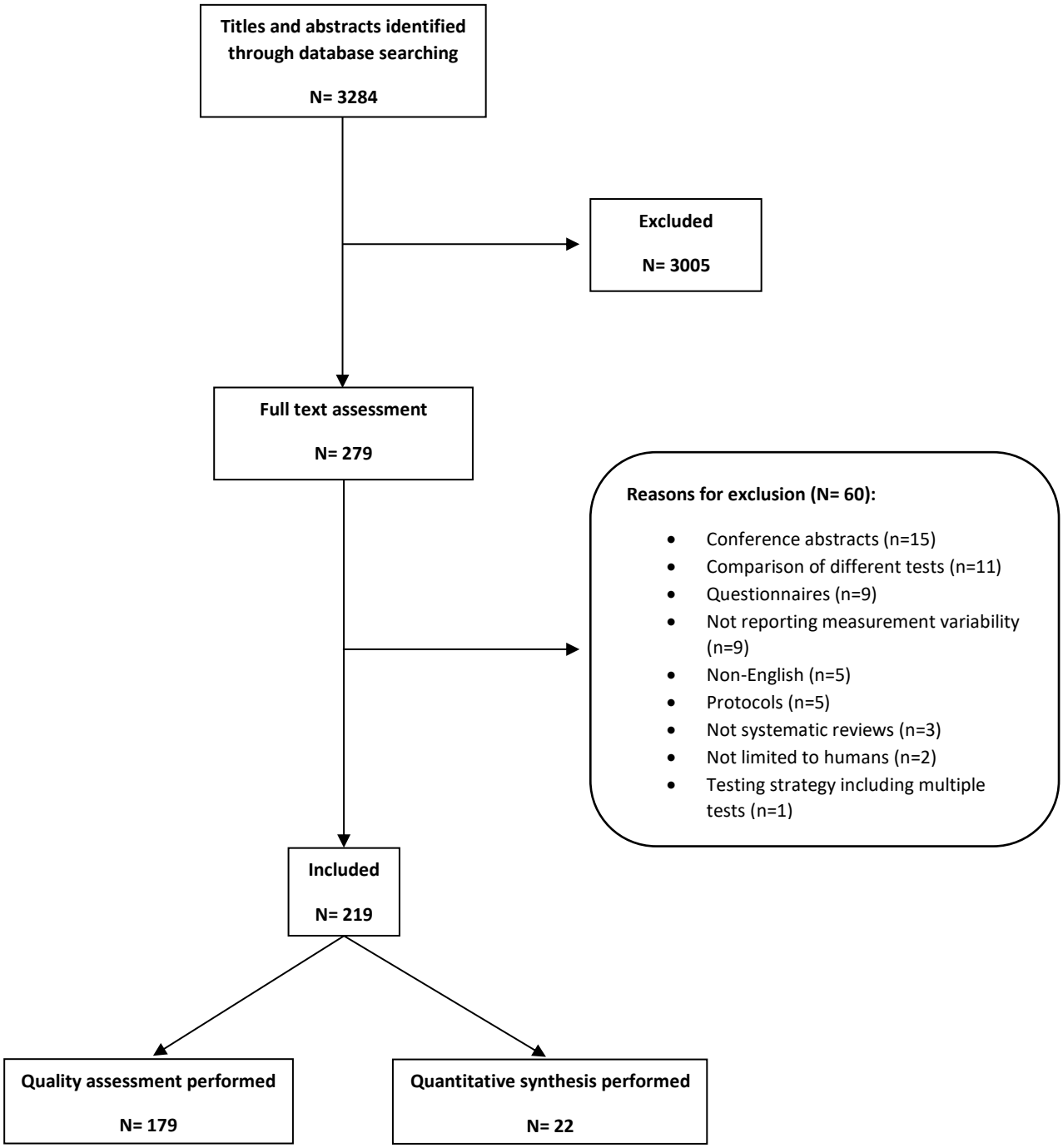


Table 4.1. Summary of the identified reviews.

Total (n=219)	Frequency (%)
Test variability as main aim (%)	163 (74)
Type of tests assessed	
Imaging[1]	62 (28)
Laboratory biomarker[2]	13 (6)
Physical performance measures[3]	114 (52)
- *Device based*	32 (15)
- *Non-device based*	56 (26)
- *Device & non-device based*	26 (12)
Physiologic[4]	12 (5)
Clinical examination[5]	8 (4)
Combination*	11 (5)
Type of variability examined	
Test-retest[6] only	70 (32)
Inter-observer only	10 (5)
Intra-observer only	5 (2)
Test-retest[6] & Inter-observer	4 (2)
Test-retest[6] & Intra-observer	3 (1)
Inter & Intra-observer	84 (38)
All three types	39 (18)
Not clear/ specified	4 (2)

[1]tests that produce detailed images of areas inside the human body (e.g., MRI, ultrasound).

[2]tests that are based on the analysis of blood/urine samples, or other substances of the human body (e.g., haemoglobin).

[3]tests that describe the physical performance of an individual (e.g., hand-held dynamometer or 6-minute walk test).

[4]tests that describe any physiologic characteristics of an individual (e.g., blood pressure).

[5]tests that are based on the clinical examination of an individual (e.g., osteopathic diagnostic palpatory test).

[6]variability of multiple within-individual measurements taken over time. Observer effect not considered.

* Device based physical function test & Non-device based physical function test & Imaging test (n=6); Physiologic test & Imaging test (n=2); Clinical examination & Imaging test (n=1); Device based physical function test & Imaging test (n=2).

4.4.2. Description of review methods

Table 4.2 shows the results for the evaluation of the review process of the identified systematic reviews, while details on each individual review are provided in Appendix B4. The majority of the included reviews provided a clear definition of the inclusion criteria (216, 99%), a clear presentation of the individual study characteristics (200, 91%), and searched at least

one electronic database in addition to PUBMED or MEDLINE (195, 89%). The full search strategy was provided in 117 reviews (54%), while 95 reviews (43%) briefly described the search terms in the text. Screening titles and abstracts was reported to have been carried out independently by two or more reviewers in 138 reviews (63%), while this number was significantly lower for data extraction (82, 37%). Quality assessment of the included primary studies was reported to have been conducted in 177 reviews (81%). Of the 177 reviews, approximately half (82, 47%) used tools developed for reliability studies (i.e., COSMIN [134] and QUAREL [136]), while 39/177 reviews (22%) used quality criteria selected by the review authors themselves. The quality assessment was carried out independently and in duplicate in 133 (75%) reviews.

Table 4.2. Assessment of the quality of the review process used in the identified reviews.

Total (n=219)	**N (%)**
Inclusion criteria presented (%)	
- Yes	216 (99)
- No	3 (1)
Study characteristics presented (%)	
- Yes	200 (91)
- Partially (brief description in text)	14 (7)
- Not reported	5 (2)
Databases searched (%)	
- PUBMED/ MEDLINE plus other	195 (89)
- PUBMED/ MEDLINE only	21 (10)
- Other	1 (<1)
- Not reported	2 (1)
Search terms described (%)	
- Yes (full strategy provided)	117 (54)
- Partially (some terms in text provided)	95 (43)
- No	7 (3)
Search period described (%)	
- Yes	149 (68)
- Partially (start or end date provided)	56 (26)
- No	14 (6)
Screening (%)	
- Independent and duplicate	138 (63)
- Single reviewer	31 (14)
- Not reported	50 (23)
PRISMA flow chart presented (%)	182 (83)

Data extraction (%)	
- Independent and duplicate	82 (37)
- Single reviewer	39 (18)
- Not reported	98 (45)
Quality assessment (%)	
- Yes	177 (81)
- No	42 (19)
If yes, how was this conducted? (%)	
- Independent and duplicate	135 (76)
- Single reviewer	10 (6)
- Not reported	29 (16)
- Not clear	3 (2)
QA tools used* (%)	
- COSMIN	56 (31)
- QUAREL	26 (15)
- QUADAS	16 (9)
- Other	52 (29)
- Author's own	39 (22)

*Some studies used multiple QA tools

4.4.3. Description of statistical estimates reported in the reviews

Table 4.3 includes the statistical estimates used to examine the reliability and measurement error of tests and biomarkers that were reported in the reviews. The intra class correlation (ICC) was the most commonly reported estimate (175, 80%), followed by the standard error of measurement (60, 27%), the Kappa coefficient (59, 27%), the coefficient of variation (55, 25%), standard correlation coefficients (52, 24%), the limits of agreement (40, 18%), and the repeatability coefficient (35, 16%). A description of each statistic has been provided in Chapter 2.

Table 4.3. Statistical parameters reported in the identified reviews.

Statistical parameter* (%)	**N (%)**
Intra class correlation	175 (80)
Standard error of measurement	60 (27)
Kappa coefficient	59 (27)
Coefficient of variation	55 (25)
Limits of agreement	40 (18)
Repeatability coefficient	35 (16)

Pearson's correlation	42 (19)
Spearman's correlation	10 (5)
Other	54 (25)
Quantitative synthesis performed (%)	22 (10)

*Some studies used multiple parameters

4.4.4. Description of pooling approaches used in the reviews

Of the 219 reviews, 197 presented a narrative synthesis, with only 22 (10%) performing a quantitative synthesis of the reported statistical estimates (see Table 4.4 for a summary). The majority of these reviews (6/22, 27%) were conducted in 2017, with the remaining studies being evenly split across 2011-2019. Of the 22 reviews, 16 (73%) used aggregated study-level data to produce a weighted average of the reported estimates, with the majority (11/16, 69%) using a random-effects model for this purpose. Two reviews (13%) used a fixed-effect model to produce weighted average, with one review stating that the choice of a fixed-effect model was due to the fact that the categorical variables in the included studies covered all possible categories in the literature. The remainder (3/16, 19%) did not specify the type of model being used. The most common statistical approaches used for examining the presence of significant between-study heterogeneity were the Cochran's Q and I-squared tests, reported in 6/16 (38%) and 8/16 (50%) reviews, respectively, with one review using both tests. In the remaining 3/15 reviews (20%) the presence of heterogeneity was not reported to have been examined in any way.

Three (14%) reviews performed individual patient data meta-analysis pooling the individual test-retest data obtained from the patients into one large single data set [28, 137, 138], while the remaining three (14%) described the distribution of the study estimates using summary statistics [139-141]. The methods identified are described in detail in sections 4.4.4.1-4.4.4.3.

4.4.4.1. Synthesis of aggregate study-level data to produce a weighted average estimate

Sixteen reviews used aggregate study-level data to produce a weighted average, with 4/16 (25%) reviews pooling more than one statistical parameter. The most common meta-analytic parameter was the intra class correlation, encountered in 7 reviews (44%). Other parameters included the coefficient of variation (4, 25%), the limits of agreement (3, 19%), the Kappa agreement coefficient (2, 13%), the standard error of measurement (2, 13%), the coefficient of repeatability (1, 6%), and the Pearson correlation coefficient (1, 6%), while 1 review (6%) pooled estimates of different types of correlation coefficients (ICC, Pearson, and Spearman).

Intra class correlation coefficient

Of the 7 reviews producing a weighted average for the intra class correlation, the majority (4, 57%) employed the method proposed by Hedges and Olkin [142]. One used the method proposed by Hunter and Schmidt [143], while the remainder 2 only stated that a random-effects model was used, with no other details provided.

Hedges and Olkin

Under this method, the estimates reported across studies are transformed to the Fisher's Z scale prior to meta-analysis using a fixed-effect or random-effects model. This transformation has been proposed by Fisher for normalising the sampling distribution of the Pearson correlation coefficient and the intraclass correlation coefficient. For the intra class correlation coefficient, the transformation is applied as follows

$$Z_{r_i} = 0.5 \ln\left(\frac{1 + (m_i - 1) r_i}{1 - r_i}\right), \quad (4.1)$$

were m_i denotes the measurements taken from each individual in the i_{th} study, and r_i is the correlation coefficient reported in the i_{th} study. The corresponding variance of the Z statistic is given by

$$Var(Z_{r_i}) = \frac{m_i}{2(n_i - 2)(m_i - 1)}, \quad (4.2)$$

were n_i is the number of individuals of the i_{th} study.

When using a fixed-effect model, a pooled estimated of the produced Fisher's Z values is obtained as

$$\overline{Z_{r_i}} = \frac{\sum_{1 \le i \le k} w_i Z_{r_i}}{\sum_{1 \le i \le k} w_i}, \quad (4.3)$$

where k is the total number of the studies included in the meta-analysis, and $w_i = (n_i - 3)$ is the weight assigned to the i_{th} study. When a random effects model is used, the calculation of the weight assigned to the study is extended to

$$w_i = \left(\frac{1}{n_i - 3} + \tau^2\right)^{-1}, \quad (4.4)$$

where τ^2 is the between-study variance, estimated using the method proposed by DerSimonian and Laird.

The variance of the pooled estimate $\overline{Z_{r_i}}$ is given by

$$Var_{\overline{Z_{r_i}}} = \frac{1}{\sum_{1 \le i \le k} w_i}, \quad (4.5)$$

with the corresponding 95% confidence interval for $\overline{Z_{r_i}}$ constructed as n be computed as

$$95\%\ CI_{\overline{Z_{r_i}}} = \overline{Z_{r_i}} \pm 1.96 \times \sqrt{Var_{\overline{Z_{r_i}}}} \quad (4.6)$$

The produced estimate and 95% confidence intervals are then reverted to the r_i metric using

$$r_i = \frac{e^{2Z_{r_i}} - 1}{e^{2Z_{r_i}} + m - 1} \quad (4.7)$$

Hunter and Schmidt

Rabelo et al [144] evaluated the reliability of dynamometer-based muscle strength measurements in post-stroke individuals with chronic hemiparesis. The authors produced a weighted average for the intra class correlation using the method developed by Hunter and

Schmidt [143]. This method was originally proposed to summarize the correlation between two variables in the field of psychometric meta-analysis (also known as validity generalisation), and emphasises the need to correct the variance in population correlations from the sampling error, as well as other sources of error ("artifacts"). These may include imperfect construct validity in both variables, the presence of error in the measurement of each variable, or samples not containing the full range of variation on both variables. Although the recommended corrections are considered the greatest strength of this method, this chapter focuses only its simplest form, which is correcting the population variance estimate from sampling error. In contrast to the method proposed by Hedges and Olkin [142], Hunter and Schmidt [143] recommend the use of the observed correlation estimates without applying any transformation. An average estimate of the correlation, weighted by sample size, is computed as

$$\overline{r_i} = \frac{\sum_{1 \le i \le k} n_i r_i}{\sum_{1 \le i \le k} n_i} \tag{4.8}$$

where k represents the total number of studies, and r_i and n_i are the correlation estimate and sample size reported in the i_{th} study, respectively. The variance in population correlations is in turn obtained by subtracting the sampling error variance from the variance in sample correlations, as

$$\sigma_{r_p}^2 = \sigma_{\overline{r_i}}^2 - \sigma_\varepsilon^2 \tag{4.9}$$

where $\sigma_{\overline{r_i}}^2$ and σ_ε^2 represent the variance in sample correlations and sampling error variance estimates, respectively. The variance in sample correlations is estimated as

$$\sigma_{\overline{r_i}}^2 = \frac{\sum_{1 \le i \le k} n_i \, (r_i - \overline{r_i})^2}{\sum_{1 \le i \le k} n_i}, \tag{4.10}$$

while the sampling error variance is estimated as

$$\sigma_{\varepsilon}^{2} = \frac{(1-\overline{r_i}^{2})^{2}}{\overline{N}-1}, \tag{4.11}$$

where $\overline{N}$ represents the average sample size. A 95% confidence interval around $\overline{r_i}$ is available through

$$95\%\ CI_{\overline{r_i}} = \overline{r_i} \pm 1.96 \times \sqrt{\sigma_{r_p}^{2}} \tag{4.12}$$

Coefficient of variation

Of the 4 reviews producing a weighted average for the coefficient of variation, 2 stated the use of a random-effects model without providing any other details on how the estimates were pooled (e.g., weight assigned to studies, estimation of the between-study variance), while the remainder did not specify the type of model being used.

Limits of agreement

Of the 3 reviews producing pooled limits of agreement, one employed the method proposed by Williamson et al [145], while the remainder only stated the use of the DerSimonian and Laird random effect model.

Method proposed by Williamson et al

Yoon et al [30] constructed pooled 95% limits of agreement to examine the inter and intra-observer variability in tumour burden measurements, produced via computed tomography and according to the Response Evaluation Criteria in Solid Tumours (RECIST) guideline. This was done using the methods presented in Williamson et al [145]. If y_{1ij} and y_{1ij} represent 2

measurements from the i_{th} patient ($i = 1, \dots, n_j$) in the j_{th} study ($j = 1, \dots, k$), taken from 2 different observers or from the same observer twice, the mean difference between the measurements (μ_{d_j}) and corresponding standard deviation for the j_{th} study (SD_{d_j}) are estimated as

$$\mu_{d_j} = \sum_{i=1}^{n_j} \frac{d_{ij}}{n_j}, \tag{4.13}$$

and

$$SD_{d_j} = \sqrt{\sum_{i=1}^{n_j} \frac{(d_{ij} - \bar{d_j})^2}{n_j - 1}}, \tag{4.14}$$

where $d_{ij} = y_{1ij} - y_{1ij}$, and n_j is the total number of individuals recruited in the j_{th} study. These are the two parameters required for the construction of the limits of agreement. For the mean differenceμ_{d_j}, the the corresponding sampling error variance of the j_{th} study is given by

$$Var\left(\mu_{d_j}\right) = \frac{(SD_{d_j})^2}{n_j}, \tag{4.15}$$

where SD_{d_j} is the standard deviation of the difference, and n_j is the sample size of the j_{th} study. When using a fixed-effect model, a weighted average estimate for μ_{d_j} is obtained as

$$\overline{\mu_{d_j}} = \frac{\sum_{1 \le j \le k} w_j \mu_{d_j}}{\sum_{1 \le j \le k} w_j} \tag{4.16}$$

where k is the total number of the studies included in the meta-analysis, and $w_j = \frac{1}{Var\left(\mu_{d_j}\right)}$ is the weight assigned to the j_{th} study. When a random effects model is used, the calculation of the weight assigned to the study is extended to

$$w_j = \frac{1}{Var\left(\mu_{d_j}\right) + \tau^2}, \tag{4.17}$$

where τ^2 is the between-study variance, estimated using the method proposed by DerSimonian and Laird.

ForSD_{d_j}, the the corresponding sampling error variance of the j_{th} study is given by

$$Var\left(SD_{d_j}\right) = \frac{(SD_{d_j})^2}{2n_j}, \tag{4.18}$$

A weighted average $\overline{SD_{d_j}}$ is then calculated in a similar fashion to $\bar{d}$. The pooled 95% limits of agreement are then constructed as

$$95\%\ LoA = \overline{\mu_{d_j}} \pm 1.96 \times \overline{SD_{d_j}} \tag{4.19}$$

Cohen's Kappa coefficient

Of the 2 reviews producing a weighted average for the Kappa coefficient, one employed the method proposed by Sun [146], and one only stated that a random-effects model was used, with no other details provided.

Method proposed by Sun

Lange et al [147] carried out a random effects meta-analysis of Cohens Kappa's as proposed by Sun [146], in order to examine the inter and intra-observer reliability of physical examination tests used for the diagnosis of shoulder pathologies. The Kappa value of the i_{th} study is estimated as

$$\kappa_i = \frac{p_{oi} - p_{ei}}{1 - p_{ei}} \tag{4.20}$$

where p_{oi} is the proportion of cases where agreement was achieved, and p_{ei} the proportion of cases where the agreement was expected by chance. For the corresponding variance, the authors provide an approximation using the calculations provided by Everitt [148], and assuming a binomial distribution for p_0. This is given by

$$Var(\kappa_i) = \frac{p_{0i}(1 - p_{0i})}{n_i(1 - p_{ei})^2}, \tag{4.21}$$

were n_i is the sample size of the i_{th} study. When using a fixed-effect model, a weighted average estimate is produced as

$$\overline{\kappa_l} = \frac{\sum_{1 \le i \le k} w_i \kappa_i}{\sum_{1 \le i \le k} w_i} \tag{4.22}$$

where k is the total number of the studies included in the meta-analysis, and $w_i = \frac{1}{Var(\kappa_i)}$ is the weight assigned to the i_{th} study. When a random effects model is used, the calculation of the weight assigned to the study is extended to

$$w_i = \frac{1}{Var(\kappa_i) + \tau^2}, \tag{4.23}$$

where τ^2 is estimated through

$$\tau^2 = \frac{1}{k-1} \sum_{1 \le i \le k} (\kappa_i - \frac{1}{k} \sum_{1 \le i \le k} \kappa_i)^2 - \frac{1}{k} \sum_{1 \le i \le k} Var(\kappa_i) \tag{4.24}$$

The variance of the pooled estimate $\overline{\kappa_l}$ is given by

$$Var_{\overline{\kappa_l}} = \frac{1}{\sum_{1 \le i \le k} w_i}, \tag{4.25}$$

with the corresponding 95% confidence interval for $\overline{\kappa_l}$ constructed as n be computed as

$$95\%\ CI_{\overline{\kappa_l}} = \overline{\kappa_l} \pm 1.96 \times \sqrt{Var_{\overline{\kappa_l}}} \tag{4.26}$$

Standard error of measurement

Of the 2 reviews producing a weighted average for the standard error of measurement, 1 stated the use of a random-effects model without providing any details on how the between-study variance was estimated, while the remaining study did not specify the type of model used.

Method used in Reavis et al

Reavis et al [149] carried out a random effects meta-analysis of standard errors of measurement (SEM) reported in studies assessing the test-retest variability of Distortion Product Otoacoustic Emission (DPOAE). In comparison to a standard random effects model (introduced in Chapter 3), the authors included an additional random effects parameter to account for the time difference between the baseline and retest measurements taken in each individual study. The authors stated that this was due to the fact that the correlation between a baseline and a repeated measurement tends to decrease over time. The study estimates and corresponding sampling variances were entered into the following model

$$SEM_i = (\beta_o + \delta_i) + (\beta_1 + \varphi_i) \times D_i + \varepsilon_i \tag{4.27}$$

where β_o represents the population mean, β_1 represents the change in SEM_i per day of follow-up, D_i represents the difference in days between the baseline and retest measurement, $\delta_i \sim N(0,\ \tau^2)$ and $\varphi_i \sim N(0,\ \gamma^2)$ represent study-specific random effects around β_o and β_1, respectively, and $\varepsilon_i \sim N(0,\ Var(SEM_i))$ represents the within-study error variance. For $Var(SEM_i)$, the authors used the formula derived by Kristof [150]. This is given by

$$Var(SEM_i) = \frac{{SEM_i}^2}{n_i}\left[n_i - \left(\frac{\Gamma(\frac{n_i+1}{2})}{\Gamma(\frac{n_i}{2})}\right)^2\right] \tag{4.28}$$

However, the authors did not provide any further details on how τ and γ were estimated.

Method used in Rozema et al

Rozema et al [151] examined the measurement error of various biometric devices used ophthalmic practice. An average estimate for the standard error of measurement was calculated as

$$SEM = \sqrt{(SD_{repeat})^2 + (SD_{reprod})^2} \tag{4.29}$$

where SD_{repeat} represents a weighted average estimate for the standard deviation of measurements performed by the same operator within the same testing session, and SD_{reprod} represents a weighted average estimate for the standard deviation of measurements performed by the same operator at different testing sessions, or by different operators. The authors state that these two average estimates, SD_{repeat} and SD_{reprod}, were weighted by the number of individuals recruited in each study. However, the authors do not specify whether a fixed-effect or a random-effects model was employed to produce the two weighted average estimates.

Coefficient of repeatability

Serai et al [152] assessed the repeatability for magnetic resonance (MR) electrography when used for measuring liver stiffness. In each included study, the repeatability of the MR electrography was examined using the percentage repeatability coefficient, defined as

$$RC = 1.96 \times \sqrt{2} \times CV_w \tag{4.30}$$

where CV_w represents the within-subject coefficient of variation, expressed as a percentage. A pooled estimate was produced as

$$\overline{RC_\iota} = \frac{\sum_{1 \le i \le k} n_i RC_i}{\sum_{1 \le i \le k} n_i} \tag{4.31}$$

where k is the total number of the studies included in the meta-analysis, and n_i is the sample size, functioning as a weight assigned to the i_{th} study. A 95% confidence interval was constructed using the bootstrap percentile method (introduced in Chapter 1). The 250th and

9750th largest estimates of $\overline{RC_i}$ produced from 10000 bootstrap samples were used as the lower and upper 95% confidence bounds, respectively.

Pearson correlation coefficient

Navarro et al [147] carried out a meta-analysis of Pearson correlation coefficients (in addition to intra class correlation coefficients) in order to examine the reproducibility of surface topography for the evaluation of the adolescent idiopathic scoliosis. For this purpose, the authors used the method proposed by Hedges and Olkin [142]. Similar to the intra class correlation, the estimates reported across studies are transformed to the Z-scale, and weighted average estimate and corresponding 95% confidence intervals are calculated using equations 4.3 and 4.6. However, when implementing this method to Pearson correlation coefficients, the transformation applied to each study estimate (see equation 4.1) reduces to $m_i = 2$, as this statistic can only be used for pairwise comparisons.

4.4.4.2. Synthesis of individual participant data

Three reviews carried out an individual patient data (IPD) meta-analysis, analysing the test-retest data obtained from each participant in each study identified in the search as one single large data set. Kramer et al [138] and Langen et al [147] examined the repeatability of quantitative fluorothymidine (^{18}F-FLT) and fluorodeoxyglucose (^{18}F-FDG) measurements produced by positron emission tomography, respectively, which are used to evaluate response to antitumor therapy. Kramer et al [138] assessed the repeatability of the ^{18}F–FLT measurements using the repeatability coefficient (RC), the R-square value, and the intra class correlation (ICC). The authors initially calculated the percentage differences between the test and retest measurements for each patient, as follows

$$\%Diff = \frac{(retest - test)}{(retest + test)/2} \times 100. \quad (4.32)$$

The repeatability coefficient was in turn calculated as $\pm 1.96 \times SD_{\overline{\%diff}}$, where $SD_{\overline{\%diff}}$ denotes the standard deviation of the percentage differences between the test and retest measurements (formula for calculating $SD_{\overline{\%diff}}$ was not provided by the authors). The R-square value was produced by regressing the retest on the test measurements. The intra class correlation was produced using a mixed effects linear model.

Langen et al [147] assessed the repeatability of the ^{18}F–FLT measurements using the repeatability coefficient (RC) and the intra class correlation (ICC). The coefficient of repeatability was calculated as 1.96 times the standard deviation of the differences between the measurements, with any differences between 2 measurements being 95% likely to be attributed to a true change rather than measurement error. The intra class correlation was produced through a random effects model, with the study and subject effects both treated as random.

Tagmouti et al [28] assessed the reproducibility of Interferon Gamma (IFN-g) Release Assays (IGRA). The variability in the repeated measurements within the same subject was then summarised using the coefficient of variation and the intra class correlation. The coefficient of variation was calculated as the standard deviation of the repeated measurements divided by the grand mean and multiplied by 100. The intra class correlation was obtained through a mixed effects model, with the subject effects incorporated into the model as random and the study effects as fixed, due to the small number of studies. A Bland-Altman plot and the Kappa agreement statistic were also used to assess the agreement between the repeated measurements. For the Kappa agreement, the test-retest values were classified according to the cut-point recommended by the manufacturer.

4.4.4.1. Synthesis of aggregate study-level data using summary statistics

Three reviews used summary statistics to describe the distribution of the reliability estimates reported in the identified studies. Powden et al [139] examined the reliability of the weight-bearing lunge test (WBLT), which measures dorsiflexion range of motion in the ankle joint. The distribution of the reported intraclass correlations and minimum detectable changes were described using the mean, median, standard error, minimum, and maximum value.

Rondoni et al [140] investigated the impact that the type of device (technological v low-cost) and the direction of movement (flexion and extension v rotation v side bending) have on the reliability of the ACROM (active cervical range of motion) measurements, in patients with non-specific neck pain. For each category, the authors calculated the mean and standard deviation of the reported intra class correlations to summarise the reliability of the ACROM measurements.

Welton et al [141] examined the reproducibility of Graph-Theoretic Brain Network Metrics on healthy humans. The authors presented the median value and the range of the reported intra class correlations in order to summarise the reliability of each metric.

Table 4.4. Summary of synthesis methods used in the identified reviews.

Study	Year	Pooled parameter	Type of synthesis	Description of method
Aarsand et al [153]	2018	Coefficient of variation	AD[1] meta-analysis	Not specified
Cavaleri et al [154]	2017	Intra class correlation	AD[1] meta-analysis	Hedges and Olkin[3] random-effects model
Chamorro et al [155]	2017	Limits of agreement	AD[1] meta-analysis	DerSimonian-Laird random-effects model
De Langen et al [137]	2012	Intra class correlation, Repeatability coefficient	IPD[2] meta-analysis	Random-effects model. Patient and study effects as random.
Gonzalez Lao et al [156]	2019	Coefficient of variation	AD[1] meta-analysis	Not specified
Hunter et al [157]	2011	Intra class correlation, Coefficient of variation, Kappa coefficient	AD[1] meta-analysis	Random-effects model (no other details provided)

Kleijn et al [158]	2012	Intra class correlation, Limits of agreement	AD[1] meta-analysis	Hedges-Olkin[3] random effects model for intra class correlations, DerSimonian-Laird random-effects model for limits of agreement
Kramer et al [138]	2018	Intra class correlation, Repeatability coefficient, R-square	IPD[2] meta-analysis	Mixed-effects model. Patient effects as random, study effects as fixed
Lange et al [147]	2017	Kappa coefficient	AD[1] meta-analysis	Sun random-effects model
Navarro et al [159]	2019	Intra class correlation, Pearson correlation	AD[1] meta-analysis	Hedges and Olkin[3] random-effects model
Powden et al [139]	2015	Intra class correlation, Repeatability coefficient	Descriptive statistics	Mean, median, standard error, minimum, maximum.
Rabelo et al [144]	2016	Intra class correlation	AD[1] meta-analysis	Hunter and Schmidt random-effects model
Reavis et al [149]	2015	Standard error of measurement	AD[1] meta-analysis	Random-effects model
Reichmann et al [160]	2011	Intra class correlation, Coefficient of variation	AD[1] meta-analysis	Random-effects model
Rondoni et al [140]	2017	Intra class correlation	Descriptive statistics	Mean, standard deviation
Rozema [151]	2014	Standard error of measurement	AD[1] meta-analysis	Not specified
Salamh et al [161]	2019	Intra class correlation	AD[1] meta-analysis	Hedges and Olkin[3] random-effects model
Serai et al [152]	2017	Repeatability coefficient	AD[1] meta-analysis	Fixed-effect model using sample size as study weight
Tagmouti et al [28]	2014	Intra class correlation	IPD[2] meta-analysis	Mixed-effects model. Patient effects as random, study effects as fixed
Weiner and McGrath [29]	2017	Intra class correlation, Pearson correlation	AD[1] meta-analysis	Hedges and Olkin[3] fixed-effect model
Welton et al [141]	2015	Intra class correlation	Descriptive statistics	Median, IQR
Yoon et al [30]	2016	Limits of agreement	AD[1] meta-analysis	DerSimonian-Laird random-effects model

[1] Aggregate study-level data, [2] Individual participant data,

[3] Fisher's Z transformation is applied to correlation coefficients prior to meta-analysis

4.5. Discussion

Evaluation of the review process

The first part of this chapter evaluated the systematic methods adopted in the identified reviews. It was encouraging to see that the majority of the reviews provided a clear description of the inclusion criteria (99%) and study characteristics (91%), and searched at least one

database in addition to PUBMED/ MEDLINE (89%), which is in line with guidelines for systematic reviews more generally [162, 163].

On the other hand, only 117 reviews (54%) provided the full search strategy, which in some cases was often too narrow and likely to be incomplete, with less than half (43%) providing a description of the search terms in the text. Screening titles and abstracts was reported to have been conducted independently by two or more reviewers in only 63% of reviews. Published evidence suggests that significantly more studies are missed when screening is performed by a single reviewer [164]. The number of reviews reporting to have carried out independent and duplicate data extraction was even lower (82, 37%), with approximately half of the reviews (98, 45%) not clearly describing how data extraction was carried out. The assessment of the methodological quality of primary studies was reported to have been conducted in 177 reviews (81%). In three quarters of these (135, 76%), the methodological quality was assessed independently by two or more authors, while 29 (16%) did not provide any information on the approach to quality assessment. A wide variety of risk of bias assessment tools was observed, with less than half (82, 46%) using tools specifically intended for examining the reliability of measurements, while 39 reviews (22%) used quality criteria selected by the review authors themselves.

Evaluation of the statistical approaches used for data synthesis

The second part of the review aimed to investigated statistical approaches used for the quantitative synthesis of the results reported across studies. Only 22/219 (10%) reviews attempted a quantitative synthesis of the reported data, with the majority (73%) performing a meta-analysis of aggregated study-level data to produce a weighted average estimate.

The most popular meta-analytic method (used in five reviews) was that proposed by Hedges and Olkin [142], where the correlation estimates are converted to the Z-scale prior to the meta-analysis. Four reviews employed this method to produce a weighted average for the

intra class correlation or the Pearson correlation coefficient, while one converted different types of correlation estimates reported across studies (ICC, Pearson, and Spearman) to a Fisher's Z-score, which was then used as a standardised common meta-analytic estimate. However, the appropriateness of the latter is questionable, and may only be performed when strong evidence of no systematic differences between the measurements is provided [15]. An alternative approach for pooling correlation coefficients included the method developed by Hunter and Schmidt [143], while methods were also identified for the limits of agreement, the standard error of measurement, and the repeatability coefficient.

All of the identified meta-analytic methods have two limitations. Firstly, all parameters are expressed as a function of the within-patient variance. However, the variance is a parameter known to have a skewed sampling distribution, particularly for particularly for small numbers of observations (in this case, studies) [145]. Thus, the assumption of underlying normality of the study-level estimates that standard meta-analysis models hold (e.g., the DerSimonian and Laird random-effects model) is likely to be violated, and performing a meta-analysis without accounting for any distributional requirements may in turn lead to a biased weighted average estimate. For the intra class and Pearson correlation coefficients, the use of the Fisher's Z transformation provides a solution to this limitation, as this transformation has shown to normalize the sampling distribution of the two parameters. However, no such approach was noted in any other method identified.

Secondly, the formulas used for estimating the sampling variance of some parameters are functions, not only of the sample size, but also of the estimate itself. Such parameters include the standard error of measurement (equation 4.28), and the standard deviation of the mean difference between two within-individual measurements, which is required for constructing the limits of agreement (equation 4.18). When the estimation of the sampling variance is dependent on the parameter estimate, the use of the inverse-variance weights (which is the standard approach for weighting studies) is no more applicable, as the weight assigned to each

study will partly depend on the magnitude of the reported estimate (and not entirely on the sample size).

Three of the six reviews that did not conduct study-level meta-analysis carried out an individual patient data (IPD) meta-analysis, pooling the test-retest measurements obtained from the participants into one single large data set. All three used the intra class correlation to assess the reliability of the measurements using a model which accounted for the within-study clustering, treating the study effect either as random or as fixed (due to the small number of studies). Additional analysis included producing Bland Altman plots, regressing the retest on the test measurements to obtain the R-square value, and using the Kappa coefficient to assess agreement between the test-retest measurements. However, these estimates were produced without accounting for the within-study clustering of the participants.

Finally, three remaining reviews described the distribution of the study estimates using summary statistics. This approach does not account for the fact that some estimates are more precise than others, as no weight is assigned to the studies. One of the three [141] stated that a meta-analysis of the ICC estimates for the various graph theory metrics was not possible due to incomplete reporting of the variances, while it would also be severely limited by the heterogeneity of the methods used in the included studies. The other two reviews [139, 140] did not justify why a formal meta-analysis of the reported estimates was not attempted.

Strengths and limitations

This review was the first to explore the current practice for conducting and reporting systematic reviews of the reliability and measurement error of biomarkers, and the current state of statistical methods available for the meta-analysis of parameters of reliability and measurement error. The findings indicated important flaws in how such reviews are conducted and reported, and how parameter estimates of reliability and measurement error reported across primary studies are combined. Two different databases were searched for the

identification of the reviews (MEDLINE and EMBASE), which is in agreement with guidelines for conducting systematic reviews in general [162, 163], and will have ensured retrieval of a comprehensive set of reviews that adequately reflect current practice.

However, this review has the following limitations. First, the study selection and data extraction was carried out by 1 reviewer only, and it is advised that these two tasks should be performed independently by at least two reviewers [133]. However, this was a methodologic review aiming to provide a comprehensive reflection of current practice, and hence did not need to be as comprehensive in study identification as it would be for, e.g., a systematic review of intervention effects or test accuracy. Furthermore, the review did not consider the adequacy of tools used for assessing the quality of the primary studies included in the identified reviews, while the overall quality of the reviews was not assessed using a formal checklist (e.g., similar to AMSTAR 2 [165], which is intended for systematic reviews of healthcare interventions). However, the latter were outside the scope of this work, which aimed to provide a general overview of current practice in this under-researched area, and primarily to identify statistical approaches for the synthesis of the data reported in primary studies. Finally, it is possible that findings of this review may be slightly outdated by the time the thesis is submitted (September 2022), given that this work was completed in 2020.

4.6. Conclusion

In this methodological review of systematic reviews examining the reliability and measurement error of biomarkers, a number of limitations in the review process and meta-analytic methods used were identified. There is scope for improvement in how such reviews are conducted and reported. In a high number of reviews, the article selection, data extraction, and quality assessment was not performed by at least two independent reviewers, while the quality of the identified primary studies was often not assessed. Furthermore, the search

strategy used, as well as the approach to screening, data extraction, and quality assessment were often not clearly reported. For the quality assessment of primary studies, a number of different risk of bias tools were used, with some reviews using quality criteria selected by the authors themselves. These findings emphasize the need for more specific guidance, both for conducting and reporting such reviews, and appraising the methodological quality of the primary studies examining the reliability and measurement error of biomarkers. Such guidance will improve the quality of reviews examining the reliability and measurement error of biomarkers, and in turn allow decision making to be based on high quality and well reported evidence.

Finally, methods for the meta-analysis of study-level data were identified for most parameters of reliability and measurement error. However, alternative methods are required for key parameters of measurement error, such as the standard error of measurement, the coefficient of variation, and the limits of agreement. These methods should focus on accounting for the non-normal distribution of the parameters, and stabilizing the sampling error variance so that the method of inverse-variance weights can be applied. Less biased and more precise average estimates will then be obtained, which will help researchers provide robust quantitative evidence on the reliability and measurement error of biomarkers using a whole body of research.

5. Proposed meta-analytic approaches for parameters of measurement error

5.1. Introduction

In the previous chapter, systematic reviews reporting the reliability and measurement error of tests were identified in the literature, in order to examine the current state of the statistical methods used for the meta-analysis of estimates of reliability and measurement error. Only 22 of the 219 reviews identified (10%) attempted a quantitative synthesis of the data reported in the primary studies, with the majority performing a meta-analysis of aggregated study-level estimates using the random-effects model proposed by DerSimonian and Laird [89]. The previous chapter discussed the potential limitations that this approach has when pooling parameters of reliability and measurement error. These include i) the violation of the normality assumption that the model holds, due to the non-normal sampling distribution of these parameters, and ii) the estimation of the within sampling variance not being independent of the parameter estimate. For i), the violation of normality assumption may lead to biased average estimates, and in turn false conclusions regarding the reliability and measurement error of biomarkers. Whilst the Fisher's Z transformation proposed by Hedges and Olkin [142] provides a potential solution to this problem when pooling Pearson or intra class correlation coefficients, no such approaches was observed for other parameters, such as the limits of agreement, the standard error of measurement, and the coefficient of variation. For ii), stabilization of the sampling error variance is required, so that its estimation is independent of the estimate of the meta-analytic parameter, which will allow the method of inverse-variance weights to be implemented.

5.2. Aim

The aim of this chapter is to present alternative methods for three key parameters of measurement error. These include the limits of agreement, the standard error of measurement, and the coefficient of variation.

5.3. Proposed meta-analytic approach for the limits of agreement

In 2017, Tipton and Shuster [166] provided a framework for producing pooled estimates for the limits of agreement. Let y_{1j} and y_{2j} represent two repeated measurements for the j_{th} individual $(j = 1, \dots, n_G)$, where n_G is the total number of individuals. Assuming that the individual differences $d_j = y_{1j} - y_{2j}$ are sampled from a normal population with mean d and standard deviation SD_d, the 95% limits of agreement are constructed as

$$LoA = \hat{d} \pm 1.96 \times \widehat{SD}_d \tag{5.1}$$

, where $\hat{d} = \sum_{j=1}^{n_G} \frac{d_j}{n_G}$ is the estimated mean difference between the repeated measurements, and $\widehat{SD}_d = \sqrt{\sum_{j=1}^{n_G} \frac{(d_j - d)^2}{n_G - 1}}$ is the estimated standard deviation of the difference. Thus, constructing a pooled interval across the identified studies requires weighted average estimates for the mean difference between two repeated measurements (d), and the standard deviation of the difference (SD_d).

Method for producing a weighted average estimate for d

For the mean difference between two repeated measurements, Tipton and Shuster [166] state that the sampling distribution of the parameter can be assumed to be normal, with $E[\hat{d}] = d$ and $Var[\hat{d}] = \frac{SD_d^2}{n_G}$. A weighted average across k studies $(k = 1, \dots, n_S)$ is produced as

$$\bar{d} = \frac{\sum_{1 \leq k \leq n_s} W_k \times (\hat{d})_k}{\sum_{1 \leq k \leq n_s} W_k} \quad (5.2)$$

, where $W_k = \frac{(n_G)_k}{(\widehat{SD_d^2})_k}$ is the weight assigned to each study, and $(\hat{d})_k, (\widehat{SD_d})_k, (n_G)_k$ denote the estimated mean difference, standard deviation of the difference, and sample size of the k_{th} study, respectively. The authors highlight the importance of the parameter estimate being independent of the sampling variance, which allows for the use of the inverse-variance weights [166].

Method for producing a weighted average estimate for SD_d

For the standard deviation of the difference, Tipton and Shuster [166] state that, assuming that d_j are normally distributed, it follows that

$$\frac{(n_G - 1)\widehat{SD_d^2}}{SD_d^2} \sim \chi^2_{(n_G - 1)} \quad (5.3)$$

, which implies that $E\left[\widehat{SD_d^2}\right] = SD_d^2$ and $Var\left[\widehat{SD_d^2}\right] = \frac{2SD_d^4}{n_G - 1}$.

The authors recommend against the use of the DerSimonian and Laird model for pooling estimates of SD_d^2 reported across studies, for two reasons. First, the sampling distribution of the parameter is not normal, particularly when the sample size is small, which violates the normality assumption that the model holds. Second, the sampling variance is a function, not only of the sample size, but also of the estimate itself. Tipton and Shuster suggest using the log-transformation as a potential solution to these two issues [166]. Following the use of the delta method (introduced in Chapter 2), the authors state that the sampling distribution of $log\widehat{SD_d^2}$ is approximately normal, with

$$E\left[log\widehat{SD_d^2}\right] = logSD_d^2 + \frac{1}{n_G - 1}. \tag{5.4}$$

Based on equation 5.4, Tipton and Shuster [166] state that the quantity $\left[log\widehat{SD_d^2} - \frac{1}{n_G - 1}\right]$ is an approximately unbiased estimate of $logSD_d^2$. The second term in equation 5.4 may in turn be omitted even for a low number of individuals n_G (e.g., ≈10).

A weighted average estimate can then computed using equation the equation labelled as 4.9 in Tipton and Shuster [166], as

$$\overline{logSD_d^2} = \frac{\sum_{1 \leq k \leq n_S} W_k \times \left[log\widehat{SD_d^2}\right]_k}{\sum_{1 \leq k \leq n_S} W_k} \tag{5.5}$$

, where $\left[log\widehat{SD_d^2}\right]_k$ is the produced estimate for the k_{th} study, and $W_k = \frac{1}{Var\left[log\widehat{SD_d^2}\right]_k}$ is the corresponding weight, with $Var\left[log\widehat{SD_d^2}\right]_k$ denoting the sampling variance of the k_{th} study. For estimating $Var\left[log\widehat{SD_d^2}\right]_k$, the authors provide the following formula

$$Var\left[log\widehat{SD_d^2}\right]_k = \frac{2}{(n_G)_k - 1} \tag{5.6}$$

, where $(n_G)_k$ is the sample size of the k_{th} study. The derivation of this formula is based on the delta method, which was described in Chapter 2.

Pooled limits of agreement and 95% confidence intervals

A pooled interval can in turn be constructed as

$$\overline{LOA} = \bar{d} \pm 1.96 \times \sqrt{e^{\overline{logSD^2}}} \tag{5.7}$$

Following the advice from Bland and Altman [47], Tipton and Shuster [166] recommend that the produced interval is presented along with the lower 95% confidence bound of the lower limit of agreement, and the upper 95% confidence bound of the upper limit of agreement [166]. These are calculated as

$$95\%CI - \overline{LOA}_L = \left(\bar{d} - 1.96 \times \sqrt{e^{\overline{logSD^2}}}\right) - t(n_{pairs} - 1, 0.025)\sqrt{\frac{3 \times e^{\overline{logSD^2}}}{n_{pairs}}} \quad (5.8)$$

, and

$$95\%CI - \overline{LOA}_U = \left(\bar{d} + 1.96 \times \sqrt{e^{\overline{logSD^2}}}\right) + t(n_{pairs} - 1, 0.025)\sqrt{\frac{3 \times e^{\overline{logSD^2}}}{n_{pairs}}} \quad (5.9)$$

, where n_{pairs} is the number of available pairs of measurements, and $t(n_{pairs} - 1, 0.025)$ is the critical value of the Student's t distribution with $n_{pairs} - 1$ degrees of freedom and 2.5% significance level.

5.4. Proposed meta-analytic approach for the standard error of measurement

Unlike the limits of agreement, no framework is available for the standard error of measurement. Thus, this chapter presents a new method producing a weighted average estimate for the standard error of measurement. The derivation of the method is based on the approach used by Tipton and Shuster for producing a weighted average estimate for SD_d (described in section 5.3).

Method for producing a weighted average estimate for SEM

The standard error of measurement is equal to the standard deviation of measurements performed within individuals. As mentioned in Chapter 2, this parameter requires the

following two assumptions: (i) the measurements produced within each individual are sampled from a normal population, and (ii) the variability in the within-individual measurements is the same across the individuals. Thus, let y_{ij} be the i_{th} measurement ($i = 1, \dots, n_I$) produced from the j_{th} individual ($j = 1, \dots, n_G$), sampled from a normal distribution with mean μ_j and variance SEM^2 (common for every individual). When using an analysis of variance (ANOVA) model, SEM^2 is estimated as

$$\widehat{SEM^2} = MS_I = \frac{SS_I}{(n_I - 1)n_G} \tag{5.10}$$

, where SS_I is the sum of squares for the within-individual variance component, n_G is the number of individuals, and n_I is the number of measurements taken from each individual. If $\bar{y}_J$ denotes the estimated mean value of the j_{th} individual, it follows that

$$SS_I = \sum_{i=1}^{n_I}\sum_{j=1}^{n_G}(y_{ij} - \bar{y}_J)^2 \Rightarrow \textit{(from equation 5.10)}$$

$$(n_I - 1)n_G\widehat{SEM^2} = \sum_{i=1}^{n_I}\sum_{j=1}^{n_G}(y_{ij} - \bar{y}_J)^2 \Rightarrow$$

$$\frac{(n_I-1)n_G\widehat{SEM^2}}{SEM^2} = \sum_{i=1}^{n_I}\left(\frac{y_{i1}-\bar{y}_1}{SEM}\right)^2 + \cdots + \sum_{i=1}^{n_I}\left(\frac{y_{in_G}-\bar{y}_{n_G}}{SEM}\right)^2$$

$$= Y_1 + \cdots + Y_{n_G}$$

Given that $\frac{y_{ij} - \mu_j}{SEM} \sim N(0,1)$ for each $j = 1, \dots, n_G$, and that $E[\bar{y}_J] = \mu_j$, each Y_j follows a chi-square distribution with $n_I - 1$ degrees of freedom [167]. From the additive property of the chi-square distribution [167], it follows that

$$\frac{(n_I - 1)n_G\widehat{SEM^2}}{SEM^2} \sim \chi^2_{n_G(n_I-1)} \tag{5.11}$$

, which implies that $E[\widehat{SEM^2}] = SEM^2$ and $Var[\widehat{SEM^2}] = \frac{2SEM^4}{n_G(n_I-1)}$.

As with $\widehat{SD_d^2}$, the use of these estimates for pooling multiple estimates reported across different studies has the following limitations. First, the sampling distribution of $\widehat{SEM^2}$ is known to be skewed to the right, and approximates normality only when the product $n_G(n_I - 1)$ is higher than 90 [168]. Second, the sampling variance is a function, not only of n_G and n_I, but also of the estimate itself.

However, similar to Tipton and Shuster [166], the log-transformation can be used in order to tackle the two limitations. When applying this transformation, the distribution of the reported study-level estimates is expected to approximate normality, with weighted average estimate computed as

$$\overline{logSEM^2} = \frac{\sum_{1 \le k \le n_S} W_k \times \widehat{logSEM_k^2}}{\sum_{1 \le k \le n_S} W_k} \tag{5.12}$$

, where $\widehat{logSEM_k^2}$ is the produced estimate for the k_{th} study, and $W_k = \frac{1}{Var\left[\widehat{logSEM_k^2}\right]}$ is the weight assigned to the k_{th} study, with $Var\left[\widehat{logSEM_k^2}\right]$ denoting the sampling variance of the k_{th} study. This can be approximated by the delta method, as follows

$$Var[log(SEM^2)] = \left(\frac{log(SEM^2)}{\partial SEM^2}\right)^2 \times \frac{2SEM^4}{n_G(n_I - 1)} = \frac{2}{n_G(n_I - 1)} \tag{5.13}$$

The corresponding variance of $\overline{logSEM^2}$ is equal to

$$Var_{\overline{logSEM^2}} = \frac{1}{\sum_{1 \le k \le n_S} W_k} \tag{5.14}$$

A 95% confidence interval for the pooled estimate can in turn be constructed as

$$95\%CI = \overline{logSEM^2} \pm 1.96 \times \sqrt{Var_{\overline{logSEM^2}}} \tag{5.15}$$

The produced weighted average and 95% confidence intervals, obtained from equations 5.12 and 5.15, can then be reverted to the original SEM scale by taking the square root of the exponential $\left(SEM = \sqrt{e^{logSEM^2}}\right)$.

5.5. Proposed meta-analytic approach for the coefficient of variation

The coefficient of variation is calculated as the ratio of the standard error of measurement (SEM) to the grand mean of the measurements (μ). Therefore, weighted average estimates for SEM and μ are required in order to derive a weighted average for the coefficient of variation.

Method for producing a weighted average estimate for SEM

For the standard error of measurement, a weighted average $\overline{SEM}$ can be obtained as described in section 5.4.

Method for producing a weighted average estimate for μ

Given that the $i\ (= 1, \dots, n_I)$ measurements produced within the j_{th} individual $(j = 1, \dots, n_G)$ are assumed normally distributed with mean μ_j and variance SEM^2, the grand mean is estimated as

$$\hat{\mu} = \frac{\overline{y_1} + \cdots + \overline{y_{n_G}}}{n_G} \tag{5.16}$$

, where $\overline{y_j}$ is the estimated mean of the j_{th} individual. Given that $\overline{y_1}, \dots, \overline{y_{n_G}}$ are independent, the corresponding sampling variance can be estimated as

$$Var[\hat{\mu}] = Var\left[\frac{\overline{y_1}+\cdots+\overline{y_{n_G}}}{n_G}\right] \Rightarrow \textit{(given that } \overline{y_1}, \dots, \overline{y_{n_G}} \textit{ are independent)}$$

$$Var[\hat{\mu}] = Var\left[\frac{\overline{y_1}}{n_G}\right] + \cdots + Var\left[\frac{\overline{y_{n_G}}}{n_G}\right] \Rightarrow \textit{(from Crawshaw and Chambers, page 437 [169])}$$

$$Var[\hat{\mu}] = \frac{Var[\overline{y_1}]}{n_G^2} + \cdots + \frac{Var[\overline{y_{n_G}}]}{n_G^2} \Rightarrow$$

$$Var[\hat{\mu}] = \frac{\widehat{SEM^2}}{n_I n_G^2} + \cdots + \frac{\widehat{SEM^2}}{n_I n_G^2} \Rightarrow$$

$$Var[\hat{\mu}] = \frac{\widehat{SEM^2}}{n_I n_G} \tag{5.17}$$

A weighted average across studies can be computed as

$$\bar{\mu} = \frac{\sum_{1 \le k \le n_s} W_k \times \widehat{\mu_k}}{\sum_{1 \le k \le n_s} W_k} \tag{5.18}$$

, where $\widehat{\mu_k}$ is the estimated grand mean of the k_{th} study, and $W_k = \frac{1}{Var[\widehat{\mu_k}]} = \frac{n_I n_G}{\widehat{SEM_k^2}}$ is the corresponding weight.

Pooled coefficient of variation and 95% confidence intervals

A weighted average for the coefficient of variation is in turn derived as

$$\overline{CV} = \frac{\overline{SEM}}{\bar{\mu}} \tag{5.19}$$

As the coefficient of variation is expressed as a ratio of two other parameters, 95% confidence intervals can be constructed via the multivariate delta method or bootstrapping, both introduced in Chapter 2.

5.6. Discussion

This chapter presents an alternative meta-analytic method for the limits of agreement, proposed by Tipton and Shuster [166], as well as two new methods developed for the standard error of measurement and the coefficient of variation. All methods focus on satisfying the assumption of underlying normality of the reported study-level estimates that models for meta-analysis hold, as well as stabilizing the sampling error variance of the parameters. The first was based on examining the sampling distribution of each parameter required for the derivation of the limits of agreement, the standard error of measurement, and the coefficient of variation. For the mean difference between two within-individual measurements and the grand mean, which are required for the derivation of the limits of agreement and the coefficient of variation, respectively, the sampling distribution is known to be normal [166, 169]. This satisfies the assumption of the underlying normality of the reported study-level estimates that standard meta-analysis models hold. However, the remaining parameters of interest are expressed as a function of the within-individual variance, which is a parameter known to have a right-skewed sampling distribution [145, 166]. The proposed methods account for potential skewness in the study-level estimates by taking the logarithm of each estimate and perform the meta-analysis using the log-transformed estimates, with the produced weighted average estimate being reverted to the original scale. This is a common approach when dealing with parameters known to have a non-symmetrical sampling distribution [170]. Furthermore, using the delta method, the estimated sampling variance is independent of the parameter estimate, which allows for the use of the inverse-variance weights. The performance of these methods is evaluated in the next chapter, where a meta-analysis of estimates reported in primary studies examining the reliability and error of grip strength measurements was carried out. Finally, as it is common that new methods are evaluated by simulation to ensure they work in the scenarios for which they were designed, the extent to which the meta-analytic methods proposed for the standard error of

measurement and the coefficient of variation are affected by different parameter inputs, different numbers of included studies, or different numbers of individuals recruited within studies should be further evaluated across different simulated scenarios.

5.7. Conclusion

This chapter presents new methods for the meta-analysis of common parameters of measurement error. The methods aim to provide a less biased and more precise average estimate of measurement error, which will allow decisions on whether a test is fit for use in medical research and practice to be based on robust summary evidence. Evaluation using case studies and simulation is required so that the performance of these methods across different scenarios is well understood.

6. Systematic Review and meta-analysis of the reliability and measurement error of hand-held dynamometers used to assess grip strength

6.1. Introduction

Grip strength (introduced in Chapter 2) is a biomarker proposed for the diagnosis, prognosis, or monitoring of numerous diseases, and is commonly measured using a handheld dynamometer. In order to be useful in clinical and research settings, the measurements of grip strength obtained by handheld dynamometers must be reliable and produced with low measurement error, so that robust conclusions can be drawn regarding changes in muscle strength. To date, a critical assessment of the evidence for the reliability and error of grip strength measurements across different populations is lacking, with the latest systematic review been published in 2011 [124]. For this purpose, a systematic review and meta-analysis were performed in order to identify primary studies examining the reliability and error of the grip strength measurements produced from handheld dynamometers. The results reported in the identified studies were then summarised using the meta-analytic methods proposed in the previous chapter.

6.2. Objectives

The primary objective was to examine the reliability and measurement error of the grip strength measurements produced from different types of handheld dynamometers (i.e., hydraulic, pneumatic mechanical, or strain [124]), across different population groups (e.g., healthy individuals, individuals with a particular disease, different genders, or different age groups). The secondary objective was to examine the effect of factors that may cause

variations in the reliability and error of grip strength measurements. Any such factors considered in this study included:

- Summary measures used per testing session (single measurement, mean of two measurements, mean of three measurements, highest of two measurements, highest of three measurements).
- Body posture during the assessment (sitting, standing).
- Different hands tested in populations with hand injuries (affected, contralateral hand).

6.3. Review methods

6.3.1. Inclusion criteria

Types of study designs

Studies of any design reporting the reproducibility of two or more measurements of maximal grip strength produced within individuals were eligible for inclusion. The measurements may be taken either from the same observer, or from multiple different observers. Any time interval between the measurements was accepted (i.e., from measurements being performed within a single testing session up to several months apart), as long as the purpose was not to determine changes in the health status of the patients between the measurement sessions. Studies assessing grip strength endurance or sustained grip strength were excluded.

Participants

Studies were restricted to those conducted in adults with any or no pre-existing condition or illness. Studies including children (aged under 18) were excluded, unless subgroup data for adults could be extracted.

Index test(s)

Evaluations of any type of dynamometer for assessing grip strength were eligible, including hydraulic, pneumatic, mechanical, or strain [124]. Studies of grip strength using an isometric strength testing unit were excluded.

Measurement properties

Studies reporting any estimate of reliability and/or measurement error were included. Estimates expected to be reported in the identified studies included the intra class correlation (ICC), the Pearson correlation coefficient (PCC), the standard error of measurement (SEM), the smallest detectable change (SDC), the coefficient of variation (CV), and the limits of agreement (LoA). These were the most common estimates reported in the systematic reviews identified in the methodological review of reliability and error in the measurement of biomarkers (Chapter 4).

6.3.2. Search strategy

Electronic searches of MEDLINE and EMBASE were undertaken by an Information Specialist on 11 April 2019 to identify relevant English language studies. No date restrictions were applied. The search strategies were informed by the COSMIN [171] search filter for identification of studies on measurement properties and are provided in Appendix C1.

6.3.3. Selection of studies

Screening of titles and abstracts retrieved by the literature searches were undertaken by a single reviewer (KT or JD) following an initial pilot of 200 records, which yielded an agreement of 90% between the two reviewers. Full text assessment of any records considered potentially eligible was undertaken independently by two reviewers, and any queries were discussed and resolved by consensus. Both screening of titles/abstracts and full test assessment were undertaken using Covidence. An adapted PRISMA (Preferred Reporting Items For Systematic Reviews And Meta-Analyses) flow-chart of study selection were included to detail the study selection process [172].

6.3.4. Data extraction

Once all relevant studies were identified, information was extracted regarding:

- the characteristics of the participants tested in each study.
- the measurement protocol that was used (type of device used, position of the arm/ hand/ wrist, number of measurements, time interval between the measurements).
- the details of the observer for each testing session (same or different observer, and level of experience).
- and the estimates of reliability and measurement error that each study reported. For reviews using a measurement unit other than kilograms (e.g., pounds or Newtons), parameter estimates expressed in the original measurement scale (e.g., SEM or LoA) rather than a proportion (e.g., ICC or CV) were converted to kilograms.

One reviewer extracted the characteristics of each study, which a second reviewer checked. A data extraction form was developed, and piloted on a random sample of five reviews prior to data extraction commencing. The items that were extracted from each study are presented in Appendix C2.

6.3.5. Assessment of methodological quality

The methodological quality of the identified studies was assessed using the following seven criteria:

(1) Evidence that the patients were stable in the time between the administrations of the tests.

(2) Evidence that the time interval between the testing sessions was appropriate.

(3) Evidence that the time interval between measurements within each session was appropriate.

(4) Evidence that the measurement conditions were the same across the testing sessions.

(5) Evidence that the professional administered the test without knowledge of other repeated measurements in the same patient.

(6) Evidence that statistical methods used for reliability were appropriate.

(7) Evidence that statistical methods used for measurement error were appropriate.

The choice of the above criteria was based on the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) risk of bias tool, which was purposely designed for studies examining the reliability and measurement error of clinical outcome measures [37]. An earlier version (precursor) to this tool was used in 56 (31%) of the reviews identified for Chapter 4. An additional criterion of this tool refers to 'professionals assigning the scores or determining the values without knowledge of the scores or values of other repeated measurements in the same patients'. This criterion was not considered applicable and was not used in this case, as measurements are produced directly from the devices. Of the criteria used, each was rated on a four-point scale (very good/ adequate/ doubtful/ inadequate). A detailed description as to how each criterion was rated is presented in

Appendix C3. The quality assessment of the included primary studies was undertaken independently by two reviewers, and any queries were discussed and resolved by consensus.

6.4. Decisions made when multiple estimates are reported within studies

It is possible that some studies may report more than one estimate of the same statistical parameter. This is because studies may consider various elements of the study design (e.g., examining both the reproducibility between two different testing sessions and within the same session), the characteristics of the participants (e.g., examining whether reproducibility varies between males and females) or observers (e.g., examining whether reproducibility varies between experienced and inexperienced clinicians), and the measurement procedure (e.g., examining whether reproducibility varies across different session summary measures used for analysis). In such cases, the following rules were applied:

Study design

Where reproducibility was reported for:

- multiple observers, estimates for the most experienced observer were used.
- multiple time points, estimates from the first testing period were used.
- measurements made within the same testing session and at different testing session, within and between-session estimates were pooled separately.
- both inter-observer (i.e., two different observers assessing each participant at two different testing sessions) and intra-observer reproducibility (i.e., same observer assessing each participant at two different sessions) of the grip strength measurements, estimates reported for the two types of reproducibility were pooled separately.

Participant characteristics

- If reproducibility estimates were reported for different subgroups of participants (e.g., males/females or diseased/healthy) rather than the overall study population, the estimates were treated as separate observations in the meta-analysis, as each estimate is produced from a different subgroup of patients and thus, provides independent information in the meta-analysis.
- For any studies of patients with hand injuries, if reproducibility estimates for both the affected and contralateral hands were reported, the estimates obtained from each hand were also treated as separate observations in the meta-analysis, under the assumption of no correlation between the measurements produced from the two hands. The assumption of no correlation is based on advice by experts that measurements taken from a healthy hand are expected to be significantly higher and more consistent, compared to measurements taken from an injured hand.
- For any other presentations of handedness, reproducibility estimates obtained from the right and dominant hand were chosen over those obtained from the left and non-dominant hand, respectively. Multiple studies have reported an approximately 10% stronger grip for the dominant hand compared to the non-dominant [173-175], while others stated this was true for the right hand, but observed no effect from hand dominance in left-handed people [176-179]. Either way, the reproducibility of the measurements is not expected to differ between the dominant and the non-dominant hand, or the right and the left hand.

Measurement procedures

Where reproducibility was reported for:

- different postures, the estimates obtained from a sitting position were used, as both the American Society of Hand Therapists [123] and the Southampton research group [124] advocate a sitting position for the participants being tested.
- different summary measures (single measurement, mean, or highest of multiple measurements), those obtained from the mean of multiple session measurements were included. Evidence suggests that this is the most consistent summary measure across sessions [180], while this approach is also recommended in the instructions provided by the American Society of Hand Therapists (mean of 3 trials) [123].
- different positions of the shoulder/elbow/forearm/wrist, the estimates obtained from the following standardised protocol were included: shoulders adducted and neutrally rotated, elbows flexed at 90°, forearms and wrists in neutral position. This protocol is recommended by the American Society of Hand Therapists [123].
- different handle positions of the dynamometer, the estimates obtained from the second handle position were included, as this position has been found to be the most consistent and is advocated for use in practice [123].
- two or more dynamometers, the estimates produced by the Jamar or any other hydraulic dynamometer were included, as both the American Society of Hand Therapists and the Southampton research group advocate the use of a hydraulic dynamometer (Jamar) [123, 124]. Estimates produced by a pneumatic dynamometer were excluded from the analysis of any reproducibility parameters estimated in the actual measurement unit (SEM, SDC, LoA), as the produced values are expressed as force per palmar surface area and cannot be converted to a unit of static force alone. However, the estimates were still included for any unit-free parameters (ICC, CV).

6.5. Statistical methods

6.5.1. Statistical model used for the meta-analysis of the reported estimates

Where possible, a weighted average for each reported parameter of reliability and measurement error was calculated using the meta-analytic methods informed by Chapters 4 and 5, and presented along with 95% confidence intervals. A 95% prediction interval representing the expected range of the values for 95% of all study populations [98-100] was also constructed, as a means of expressing the between-study heterogeneity.

The DerSimonian and Laird random effects model [90] was employed in order to account for the sampling error of each study, as well as any potential heterogeneity due to different characteristics across studies, such as different study populations or different time intervals used between testing sessions. The model is expressed as

$$y_i = \mu_i + \varepsilon_i = (\mu + \delta_i) + \varepsilon_i, \tag{6.1}$$

where y_i represents the observed estimate in the i_{th} study, μ represents the overall population mean, $\delta_i \sim N(0,\ \tau^2)$ represents the difference between the mean of the population the i_{th} study was sampled (μ_i) and the overall population mean (μ), and $\varepsilon_i \sim N(0, \sigma_i^2)$ represents the estimation error due to within-study sampling variability. A weighted average was computed as

$$\bar{y} = \frac{\sum_{1 \le i \le k} w_i y_i}{\sum_{1 \le i \le k} w_i} \tag{6.2}$$

with the variance of the pooled estimate equal to

$$Var_{\bar{y}} = \frac{1}{\sum_{1 \le i \le k} w_i}, \tag{6.3}$$

where k is the total number of the studies included in the meta-analysis, and $w_i = \frac{1}{Var_{y_i} + \tau^2}$ is the weight assigned to each study, with Var_{y_i} denoting the sampling variance of the i_{th} study, and τ^2 denoting the variance between studies.

A 95% confidence interval for the pooled estimate was obtained as

$$95\%\ CI = \bar{y}\ \pm 1.96 \times \sqrt{Var_{\bar{y}}} \tag{6.4}$$

The prediction interval was constructed as

$$95\%\ PI = \bar{y} \pm t_{k-2} \times \sqrt{Var_{\bar{y}} + \tau^2} \tag{6.5}$$

where t_{k-2} is the 97.5th percentile of the t distribution with $k - 2$ degrees of freedom. The standard approach of using a t-distribution, following Riley et al [98-100], was employed, as in contrast to the Z-distribution, this distribution accounts for the uncertainty of the between-study variance (τ^2).

6.5.2. Approach used for the intra class correlation coefficient

The intra class correlation (or reliability coefficient) denotes the proportion of the total variability that is attributable to true differences between participants. A Fisher's Z transformation was applied to the intra class correlation coefficients reported in the identified studies prior to the meta-analysis, as proposed by Hedges and Olkin [142]. The formula for converting each intra class correlation to a Z value is

$$Z_r = 0.5 \ln(\frac{1+(m-1)r}{1-r}), \tag{6.6}$$

where r is the intra class correlation coefficient and m is the number of measurements produced from each participant. The corresponding variance of the above statistic is

$$Var_{Z_r} = \frac{m}{2(n-2)(m-1)}, \tag{6.7}$$

where n is the number of participants. A weighted average estimate $\overline{Z_r}$ along with a 95% confidence interval were computed using equations 6.2 and 6.4, while a 95% prediction interval was constructed using equation 6.5. The produced weighted average estimate, 95% confidence intervals, and 95% prediction intervals were then reverted to the ICC metric using

$$r = \frac{e^{2Z_r} - 1}{e^{2Z_r} + m - 1} \tag{6.8}$$

6.5.3. Approach used for the Pearson correlation coefficient

Similar to the intra class correlation, the Pearson correlation coefficients reported across studies were Z-transformed prior to the meta-analysis, as proposed by Hedges and Olkin [142]. In order to produce a weighted average estimate, a 95% confidence interval, and a 95% prediction interval, formulas 6.6-6.8 were reduced to $m = 2$, as Pearson correlation coefficients can only be used for pairwise comparisons [55].

6.5.4. Approach used for the standard error of measurement

Prior to the meta-analysis, the estimates of the standard error of measurement reported in the identified studies were initially squared and then log-transformed ($logSEM^2$), as described in Chapter 5. The corresponding variance was given by

$$Var_{logSEM^2} = \frac{2}{n \times (m - 1)} \tag{6.9}$$

where n is the number of participants, and m is the number of measurements produced from each participant. A weighted average estimate $\overline{logSEM^2}$ along with a 95% confidence interval and 95% prediction interval were produced using formulas 6.2, 6.4, and 6.5. The produced

estimates were in turn reverted to the original scale by taking the square root of the exponential $\left(\sqrt{e^{\overline{logSEM^2}}}\right)$.

6.5.5. Approach used for the smallest detectable change

A weighted average estimate for the smallest detectable change was then produced using the following formula

$$\overline{SDC} = \sqrt{2} \times 1.96 \times \overline{SEM} \tag{6.10}$$

where $\overline{SEM}$ is the weighted average for the standard error of measurement, which was calculated as described under section 6.5.4.

6.5.6. Approach used for the limits of agreement

The limits of agreement are expressed as the interval of 1.96 times the standard deviation of the differences between two repeated measurements, either side of the mean difference [66]. In mathematical notation this is expressed as

$$LoA = d \pm 1.96 \times SD_d \tag{6.11}$$

Thus, the two parameters required for producing a pooled interval across the identified studies include the mean difference between two repeated measurements (d) and the standard deviation of the mean difference (SD_d). For the mean difference d, a weighted average was produced using formula 6.2, with the sampling variance of each study-level estimate calculated as

$$Var_d = \frac{SD_d^2}{n} \tag{6.12}$$

For SD_d, prior to the meta-analysis the individual estimates reported in the identified studies were initially squared and then log-transformed ($logSD_d^2$), as described in Tipton and Shuster [166]. The corresponding sampling variance is given by

$$Var_{logSD_d^2} = \frac{2}{(n\ -\ 1)} \tag{6.13}$$

where n represents the study sample size. The produced weighted average $\overline{logSD_d^2}$ was then reverted to the original scale by taking the square root of the exponential $\left(\sqrt{e^{\overline{logSD_d^2}}}\right)$. The pooled limits of agreement were then constructed as

$$\overline{LoA} = \bar{d} \pm 1.96 \times \overline{SD_d} \tag{6.14}$$

where $\bar{d}$ and $\overline{SD_d}$ represent the weighted average estimates for the mean difference (d) and the standard deviation of the mean difference (SD_d), respectively. A 95% confidence interval for $\overline{LoA}$ was constructed as

$$95\%CI - \overline{LOA}_L = \left(\bar{d} - 1.96 \times \sqrt{e^{\overline{logSD_d^2}}}\right) - t(n-1, 0.025)\sqrt{\frac{3 \times e^{\overline{logSD_d^2}}}{n}}, \tag{6.15}$$

and

$$95\%CI - \overline{LOA}_U = \left(\bar{d} + 1.96 \times \sqrt{e^{\overline{logSD_d^2}}}\right) + t(n-1, 0.025)\sqrt{\frac{3 \times e^{\overline{logSD_d^2}}}{n}}, \tag{6.16}$$

where $t(n-1, 0.025)$ is the critical value of the Student's t distribution with $n-1$ degrees of freedom and 2.5% significance level.

6.5.7. Approach used for the coefficient of variation

The coefficient of variation is expressed as the ratio of the variability within individuals to the grand mean of the measurements, multiplied by 100 for expressing it as a percentage. In mathematical notation this is expressed as

$$CV = \frac{SEM}{\mu} \times 100 \tag{6.17}$$

Thus, the two parameters required for producing a weighted average for the coefficient of variation include the standard error of measurement (SEM) and the grand mean (μ). If individual studies were reporting the mean of two different testing sessions separately, the grand mean was calculated as

$$\mu = \frac{\mu_1 + \mu_2}{2}, \tag{6.18}$$

where μ_1 and μ_2 represent the mean values of two different testing sessions. The sampling variance of the grand mean was calculated as

$$Var_\mu = \frac{SEM^2}{n \times m} \tag{6.19}$$

A weighted average estimate for the grand mean $(\bar{\mu})$ was then produced using formula 6.2, while a weighted average estimate for the standard error of measurement $(\overline{SEM})$ is produced as described under section 6.5.4. A weighted average estimate for the coefficient of variation was in turn calculated as

$$\overline{CV} = \frac{\overline{SEM}}{\bar{\mu}} \times 100 \tag{6.20}$$

where $\overline{SEM}$ and $\bar{\mu}$ represent the weighted averages for the standard error of measurement and the grand mean, respectively. A 95% confidence interval for $\overline{CV}$ was constructed using

2.5th and 97.5th percentiles of 1000 bootstrapped samples, with bias-correction applied in case the original estimate did not lie at the 50th percentile [79].

6.5.8. Subgroup analysis

If possible, a subgroup analysis was also performed for the following factors:

- Summary measures per session used for calculating reproducibility (single measurement, mean of two measurements, mean of three measurements, highest of two measurements, highest of three measurements).
- Body posture during the assessment (sitting, standing).
- Different hands tested in populations with hand injuries (affected, contralateral hand).

For the first two factors, the assessment of grip strength based on a sitting position and the mean of three session measurements are considered the most reliable and are recommended for use in practice [123, 124]. However, this hypothesis was further tested in a subgroup analysis, as there is controversy in the literature with respect to the above recommendations.

For the summary measure used at each testing session, Hamilton et al [181] found similar test–retest reliability between the mean of three measurements and the mean of two, the maximum of three, or even a single session measurement. Furthermore, Coldham et al [182] reported that a single measurement was as reliable as the mean of three measurements, as well as less tiring for the participants.

With respect to the posture of the participants during the assessment, Shechtman et al [179] found similar test–retest reliability between a sitting and a standing testing position, for both the right and the left hand.

An additional subgroup analysis was also carried out for studies reporting reproducibility estimates for both the affected and contralateral hand. High reproducibility is required for

both hands, as the American Society of Hand Therapists recommends comparing the measurements of the two hands when estimating loss of grip strength [123].

For each subgroup, a weighted average estimate was produced using the statistical methods described under 6.5.1. The produced estimates were then compared across the different subgroups within each of the three factors. In the case of multiple estimates reported within studies, all estimates were included in the subgroup analysis performed for each potential factor of heterogeneity.

6.6. Results

6.6.1. Summary of identified studies

A total of 7130 unique records were retrieved, of which 172 were selected for full text assessment. Of the 172 records, 80 met the eligibility criteria (Figure 6.1). A summary of the included studies is presented in Table 6.1, while the characteristics of each individual study are presented in Appendices C4, C5, and C6. A wide range of populations was covered, from patients with chronic diseases such as stroke (4, 5%), advanced cancer (1, 1.3%), and end-stage renal disease (1, 1.3%), with the largest group representing healthy populations (27, 33.8%). The median age (in years) of the participants was 45.8 [35.0, 66.5], while the median proportion of male participants across studies was 48.5% [33.3%, 60.6%]. The median sample size of the studies was 35 participants [Q1=25, Q3=76]. Seventy-two studies (90.0%) examined the reproducibility of the measurements made across two or more testing sessions, with the median time interval between testing sessions being 7 days [Q1=2, Q3=7]. Of the 72, 56 (77.8%) examined the intra-observer reproducibility of the measurements alone, 10 (13.9%) considered the inter-observer reproducibility of the measurements alone, while the remaining 6 (8.3%) examined both types of reproducibility. Twelve studies (15.0%) examined the reproducibility of successive measurements made at a single testing session, either alone (n=8)

or in addition to the between-session reproducibility (n=4), with the median time interval between measurements made within the same session being 30 seconds [Q1=15, Q3=60]. Seventy-two studies (90.0%) enrolled the participants prospectively, while the remainder (8/80, 10.0%) used retrospective data from already conducted studies.

A number of different dynamometers were used; the Jamar was the most commonly reported (46/80, 57.5%), followed by the Takei (6/80, 7.5%) and Baseline (6/80, 7.5%). Half (41, 51.3%) of studies reported calibration of the dynamometer. The handle position of the dynamometer was stated in 43 studies (53.8%). Most placed the handle at the second position (27/43, 62.8%), while 11 studies (25.6%) mentioned variations across patients with respect to the handle position used. In 10/11 studies (90.9%), the variations were dependant on the hand of the participant, while in the remaining study [183] adjustments were based on the gender (2nd position used for women, 3rd position used for men).

Seventy-three studies (91.3%) described the posture of the individual, with the majority of studies testing the participants in a sitting position (62/73, 84.9%). The position of the shoulder, the elbow, and the wrist was stated in 53, 62, and 54 studies, respectively. Both hands were tested in the majority of studies (51/80, 63.8%), with different presentations of handedness including right & left (n=27), dominant & non-dominant (n=13), affected & contralateral (n=8), more affected & less affected (n=3). Half of the studies (40/80, 50%) used the mean three consecutive measurements as a summary measure per session, followed by the highest of three consecutive measurements (22/80, 27.5%). Preparatory instructions and encouragement during the procedure were reported to have been provided to the participants in 57 (71.3%) and 34 (42.5%) studies, respectively.

Figure 6.1. PRISMA flow chart.

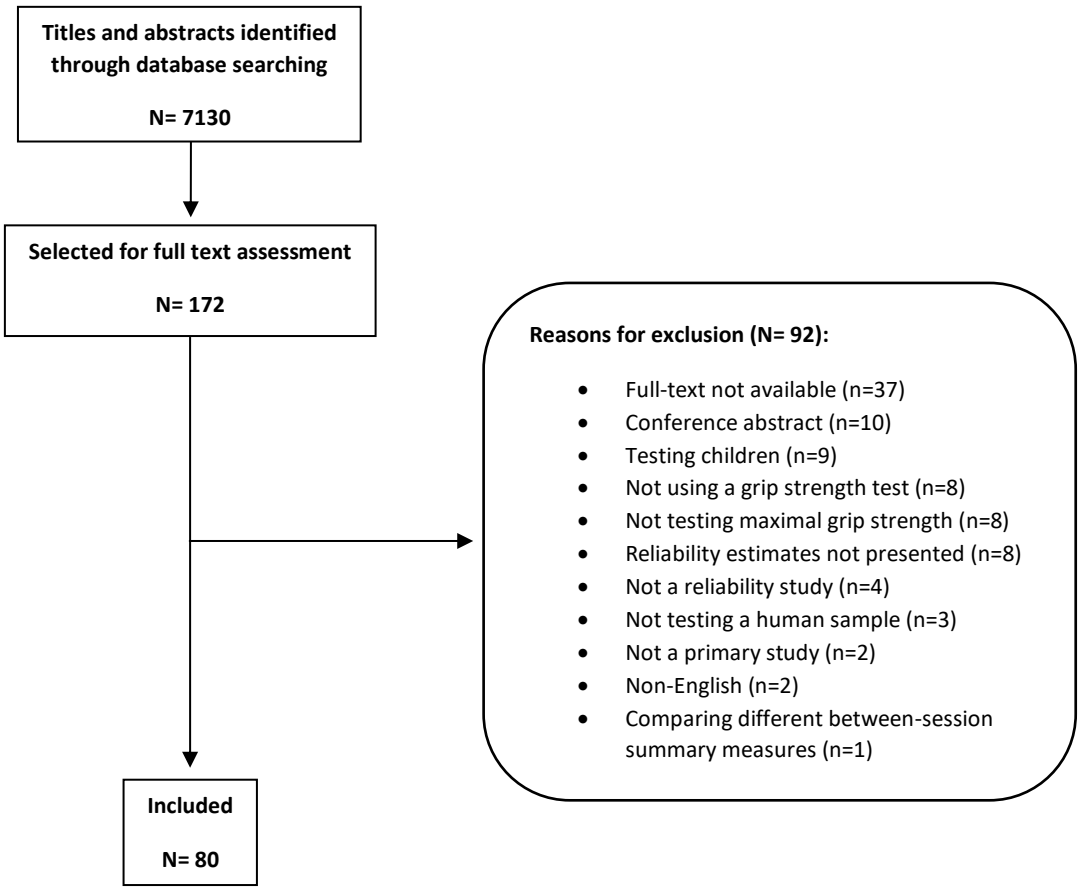


Table 6.1. Summary of the identified studies (N=80).

Study characteristics	
Design	
- Repeated measures	72 (87%)
- Cross-sectional single-session	8 (13%)
Sample size [median, IQR]	35 [25, 76]
Data collection	
- Prospective	72 (87%)
- Retrospective	8 (13%)
Type of reliability examined[1]	
- Intra-observer	56 (78%)
- Inter-observer	10 (14%)
- Both	6 (8%)
Number of testing sessions	
- One	7 (9%)
- Two	67 (84%)
- Three	3 (4%)
- Four	2 (3%)

- Ten	1 (1%)
Time interval between testing sessions (days) [median, IQR]	7 [2, 7]
Missing	12
Time interval between measurements made within testing sessions (seconds) [median, IQR]	30 [15, 60]
Missing	20
Observer characteristics	
Number of participating observers	
- One	64 (80%)
- Two	14 (17%)
- Three	2 (3%)
Where the observers experienced?	
- Yes	22 (28%)
- No	6 (8%)
- Mixture	2 (3%)
- Not reported	50 (63%)
Participant characteristics	
Medical condition	
- None (healthy participants)	27 (34%)
- Stroke	4 (5%)
- Burn injuries	3 (4%)
- Hand/wrist injuries	3 (4%)
- Dementia	3 (4%)
- Mental retardation	2 (3%)
- Charcot-Marie-Tooth	2 (3%)
- Other	26 (33%)
- Combination	10 (13%)
Age of participants (years) [median, IQR]	45.8 [35.0, 66.5]
Missing	2
Gender of participants (%) [median, IQR]	48.5 [33.3, 60.6]
Missing	4
Measurement conditions	
Device calibrated	
- Yes	41 (51%)
- Not reported	39 (49%)
Preparatory instructions provided to participants	
- Yes	57 (71%)
- No	2 (3%)
- Not reported	21 (26%)
Vocal encouragement provided	
- Yes	34 (43%)
- No	8 (13%)
- Not reported	38 (48%)
Measurement protocol	
Summary measure per session[2,3]	
- Mean of 3 measurements	40 (50%)
- Highest of 3 measurements	22 (28%)
- Mean of 2 measurements	9 (11%)
- First session measurement	8 (10%)
- Highest of 2 measurements	5 (6%)

- Single measurement	4 (5%)
- Mean of 10 measurements	2 (3%)
- Other	4 (5%)
- Not reported	3 (4%)
Tested hand(s)[3]	
- Both[4]	51 (64%)
- Dominant only	13 (16%)
- Non-dominant only	4 (5%)
- Affected only	3 (4%)
- Right only	2 (3%)
- Not reported	10 (13%)
Dynamometer used[5]	
- Jamar	46 (58%)
- Takei	6 (8%)
- Baseline	6 (8%)
- SAEHAN	3 (4%)
- Biometrics	3 (4%)
- GRIP-it	3 (4%)
- Eval Solosystem	2 (3%)
- Smedley	2 (3%)
- Other	16 (20%)
- Not reported	4 (5%)
Body posture[6]	
- Sitting position	61 (76%)
- Standing position	12 (15%)
- Sitting or lying	1 (1%)
- Not reported	7 (9%)
Handle position[7]	
- Second	27 (34%)
- Adjusted	11 (14%)
- Third	4 (5%)
- Fourth	2 (3%)
- Fifth	1 (1%)
- Not reported	37 (93%)

[1] Only applicable for repeated measures study designs

[2] 12 studies reported estimates from multiple session summary measures

[3] 3 studies reported estimates from 2 different subgroups of patients

[4] Different presentations of handedness include: right & left (n=27), dominant & non-dominant (n=13), affected & contralateral (n=8), more affected & less affected (n=3)

[5] 10 studies reported estimates from multiple dynamometers

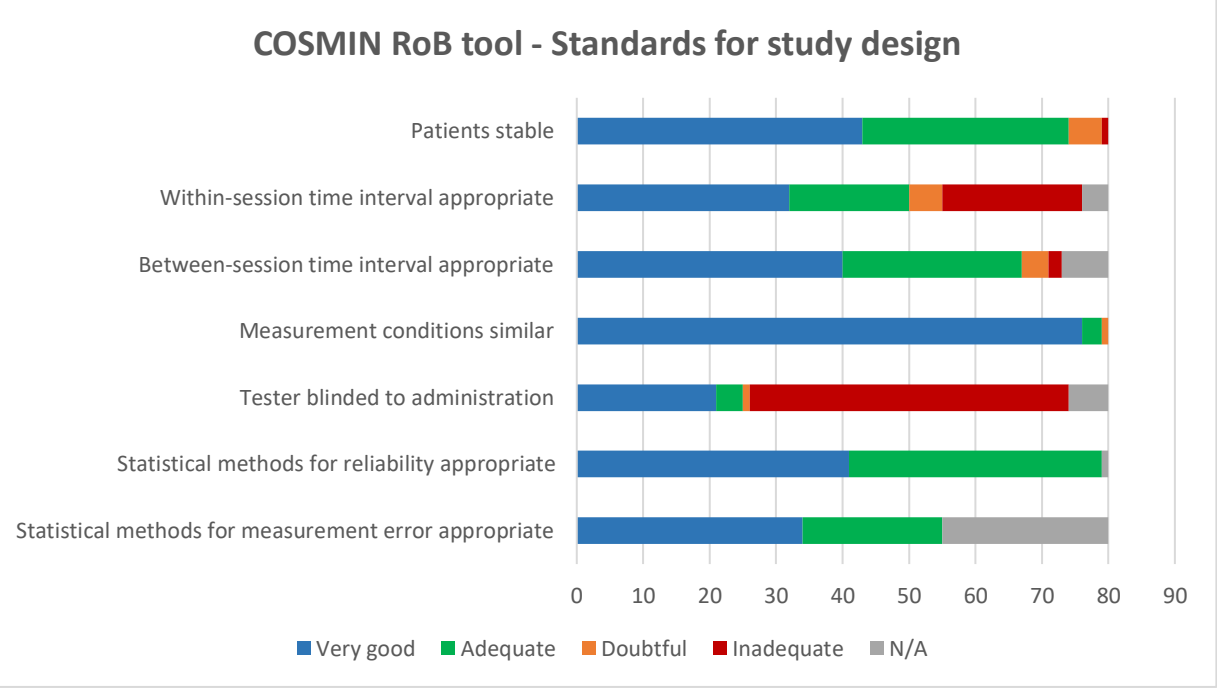
[6] 1 study reported estimates from both a sitting and a standing position

[7] 2 studies reported estimates from multiple handle positions

6.6.2. Assessment of methodological quality

The overall ratings for the seven criteria are depicted in Figure 6.2, while the ratings for each individual study are presented in Appendix C7. For the standard on blinding of the measurements, the majority of studies (49/80, 61%) provided doubtful (n=1) or inadequate (n=48) evidence that the professionals administering the test were blinded to other repeated measurements made in the same patient. For the time interval between measurements made within the same testing session, 26 of the 80 studies (33%) provided doubtful (n=5) or inadequate (n=21) evidence of an appropriate time interval. For the remaining five criteria, the ratings were at least adequate in ≥90% of the studies.

Figure 6.2. COSMIN Risk of Bias tool – Standards for study design.



6.6.3. Reproducibility of measurements taken at different testing sessions

6.6.3.1. Measurements performed from a single observer

The majority of the identified studies (62/80, 77.5%) examined the intra-observer reproducibility of the grip strength measurements across multiple testing sessions. The results are presented in Table 6.2 and Figures 6.3-6.10.

Intra class correlation

The intra class correlation was reported in 54 of the 62 studies (87.1%). The majority (50/54, 92.7%) examined the reproducibility of grip strength measurements made at two different testing sessions, with 68 observations included in the meta-analysis. The distribution of the 68 ICC values reported across studies was highly skewed [median (Q1, Q3) = 0.96 (0.94, 0.98); min= 0.42; max> 0.99], with the Z-transformation being effective in normalising the reported ICC values (see Figure 6.3). The heterogeneity across the produced Z-values appeared high (Figure 6.4), with the lower and upper quartiles of the observed distribution being equal to 1.74 and 2.30 (min= 0.44; max= 3.11). The produced weighted average Z value was 1.95 [(95% CI: 1.84, 2.06); 95% PI: (1.09, 2.80)]. The corresponding weighted average for the intra class correlation was 0.96 (95% CI: 0.95, 0.97), with a 95% prediction interval (PI) of 0.80 to 0.99.

Two studies [184, 185] examined the reproducibility of the grip strength measurements over three testing sessions, with the ICC values equal to 0.92 and 0.94 respectively. One study [186] examined the reproducibility of grip strength measurements taken at four testing sessions (ICC=0.96), while 1 additional study [187] examined the reproducibility of the grip strength measurements over 10 testing sessions, with the ICC values reported for males and females being 0.91 and 0.94 respectively.

Figure 6.3. Distribution of reported intra class correlations and the corresponding Fisher's Z-values.

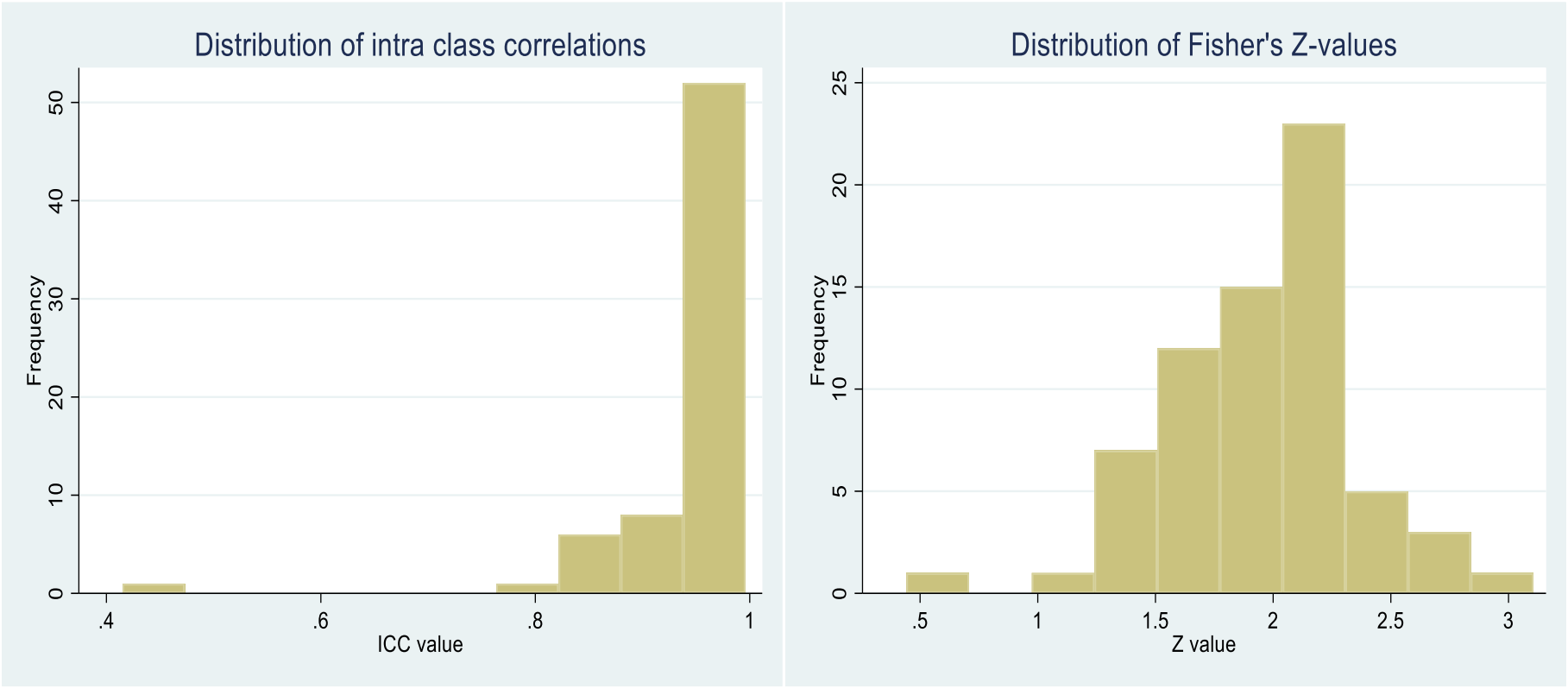
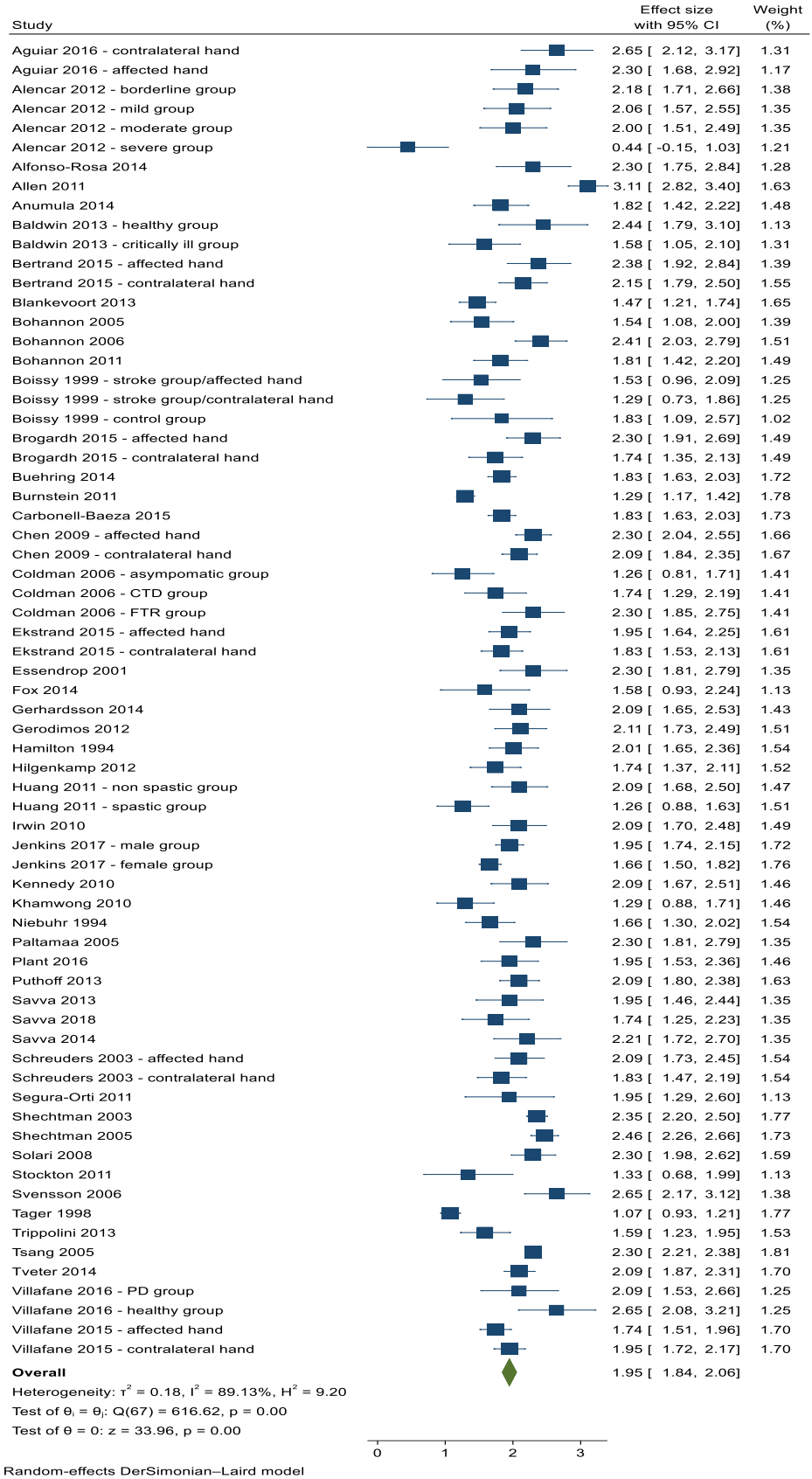


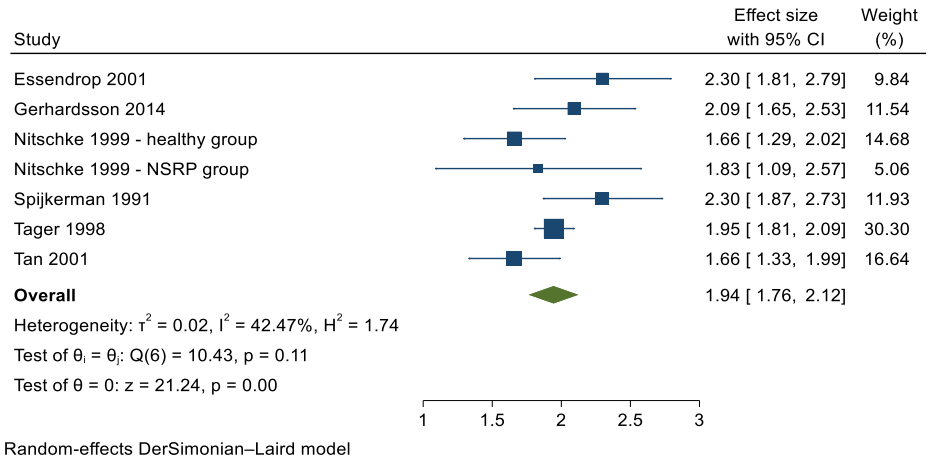
Figure 6.4. Forest plot of Fisher's Z (for intra class correlations) including studies examining intra-observer reproducibility across two different testing sessions.



Pearson class correlation

The Pearson correlation coefficient was reported in 6/62 studies (9.7%), either alone (n=3) or in addition to the intra class correlation (n=3), with 7 observations included in the meta-analysis. The produced weighted average Z-value was 1.94 [95% CI: (1.76, 2.12); 95% PI: (1.49, 2.40)]. Using formula 6.7, the corresponding Pearson correlation coefficient was 0.96 [95% CI: (0.94, 0.97)], with a 95% prediction interval of 0.90 to 0.98.

Figure 6.5. Forest plot of Fisher's Z (for Pearson correlations) including studies examining intra-observer reproducibility across two different testing sessions.



Standard error of measurement and smallest detectable change

The standard error of measurement (SEM) was reported in 33/62 studies (53.2%), with the calculation also possible in an additional four studies through the reported standard deviation of the mean difference between two testing sessions ($SEM_{consistency} = \frac{SD_d}{\sqrt{2}}$) [15], and two additional studies through the reported smallest detectable change ($SDC = \sqrt{2} \times 1.96 \times SEM$) [51]. This gives a total of 39 studies, all examining the reproducibility of the measurements across two testing sessions, with 54 observations included in the meta-analysis.

The distribution of the SEM values reported across studies was positively skewed [median (Q1, Q3) = 1.81 (1.46, 2.55); min= 0.18; max= 6.73], with the transformed values fitting the normal distribution better compared to the original SEM values (see Figure 6.6). The heterogeneity across the transformed values was high (Figure 6.7), with the lower and upper quartiles of the distribution being equal to 0.76 and 1.87 (min= -3.39; max= 3.81). The weighted average produced for the transformed values was 1.16 [95% CI: (0.92, 1.39); 95% PI: (-0.53, 2.84)].

The corresponding weighted average estimates for the standard error of measurement (SEM) and smallest detectable change (SDC) were 1.78kgs [95% CI: (1.59kgs, 2.00kgs); 95% PI: (0.77kgs, 4.14kgs)] and 4.93kgs [95% CI: (4.40kgs, 5.54kgs); 95% PI: (2.13kgs, 11.47kgs)], respectively.

Figure 6.6. Distribution of values reported for the standard error of measurement (SEM) and corresponding logarithm of squared values (log(SEM^2)).

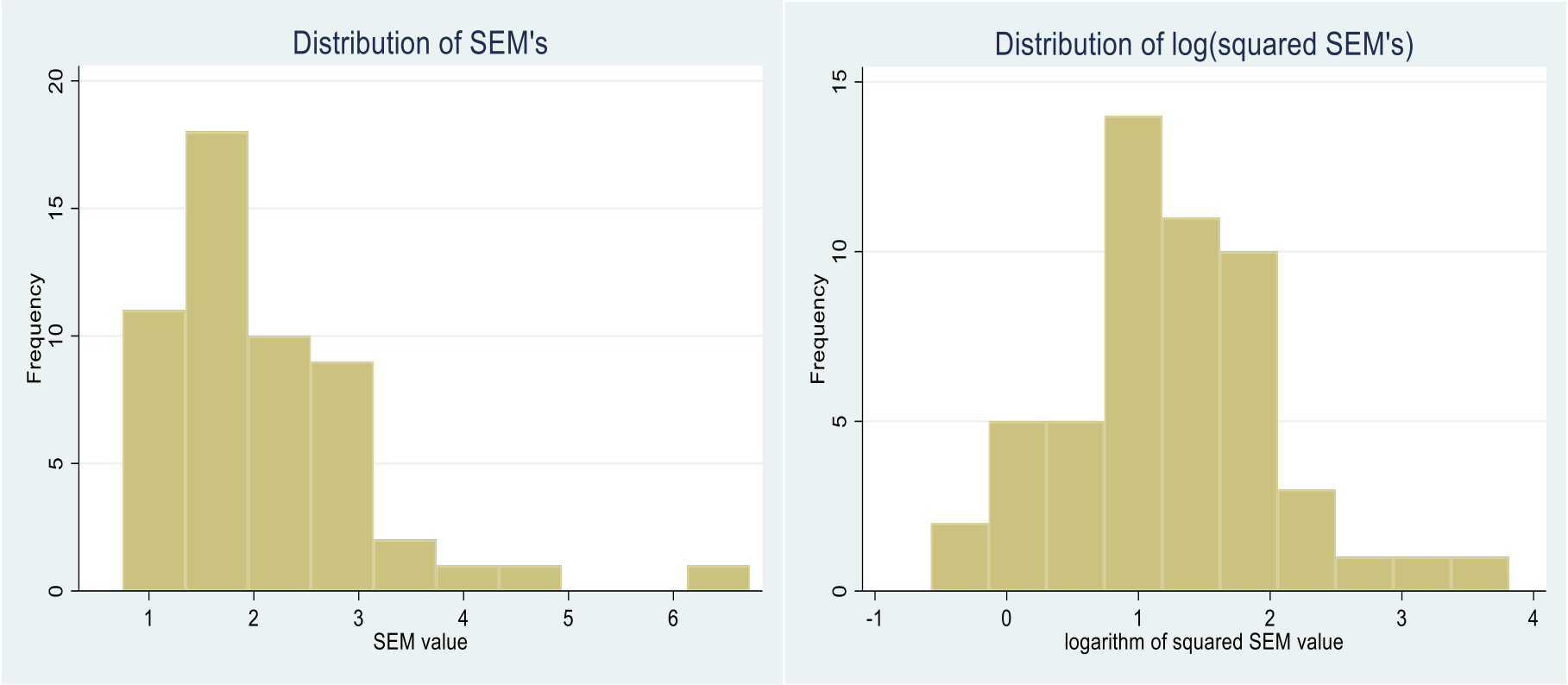
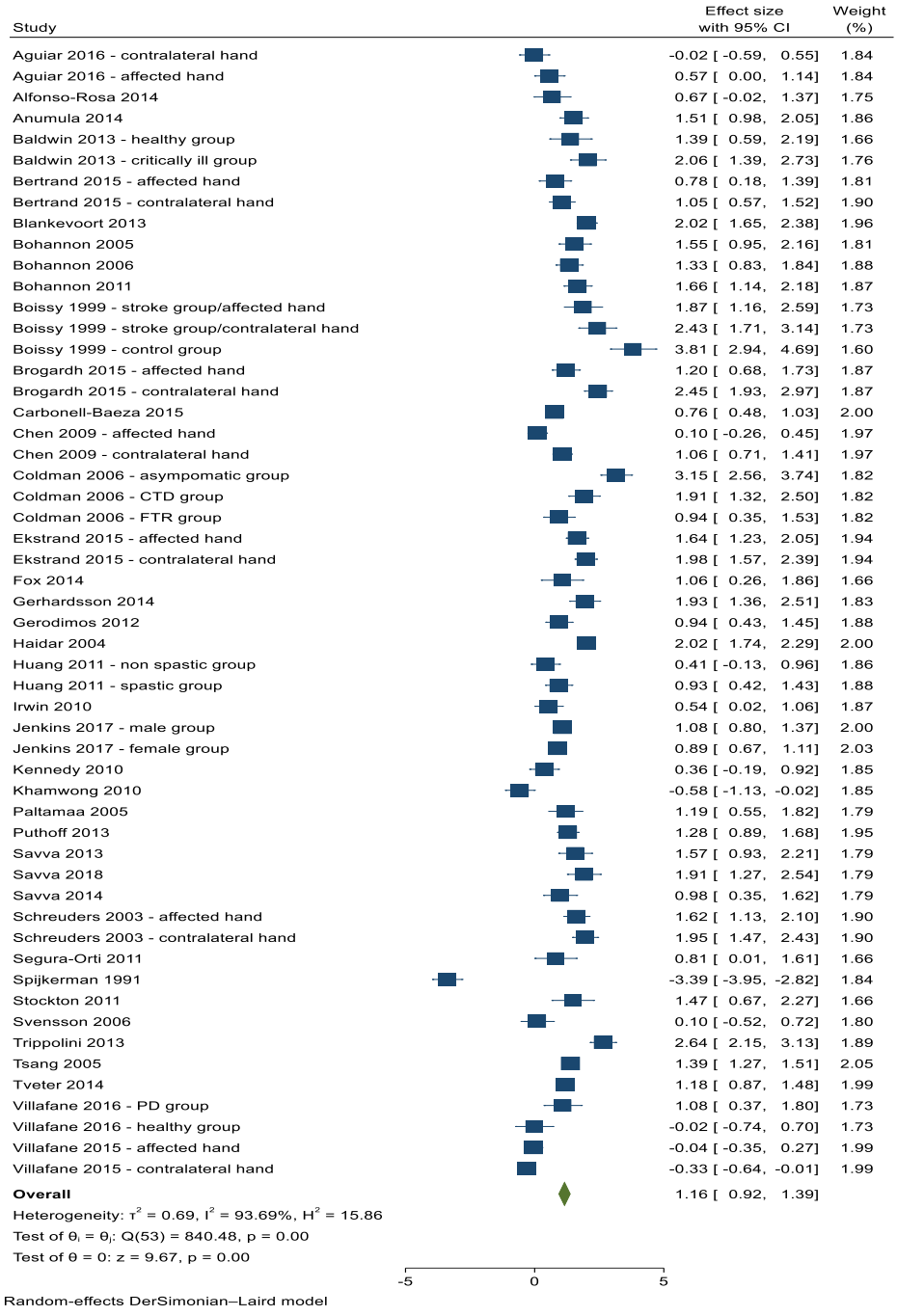


Figure 6.7. Forest plot of logSEM^2 including studies examining intra-observer reproducibility across two different testing sessions.



Limits of agreement

The limits of agreement were reported in 17/62 (27.4%) studies, with 21 observations included in the meta-analysis for each parameter required for the calculation of the two limits. For the standard deviation of the difference (SD_d), the distribution of the study values was not normal [median (Q1, Q3) = 2.94 (2.31, 3.73); min= 0.35; max= 9.99], with the applied transformation leading to a good approximation of the normal distribution (Figure 6.8). The heterogeneity across the transformed values appeared high (see Figure 6.9), with the lower and upper quartiles of the distribution being equal to 1.67 and 2.63 (min= -2.10; max= 4.61). In contrast, except for a single outlier, the study values reported for the mean difference between the two testing sessions (d) were homogenous (Figure 6.9).

The produced weighted average for the mean difference between the two testing sessions (d) and the standard deviation of the difference (SD_d) was 0.34 [95% CI: (0.17, 0.52)] and 2.57 [95% CI: (2.01, 3.28)], respectively. Using formula 6.14, the corresponding limits of agreement were 0.34 ± 1.96×2.57 = (-4.69, 5.38) [95% CI: (-6.83, 7.51)].

Figure 6.8. Distribution of values reported for the standard deviation of difference (SD_d) and corresponding logarithm of squared values ($\log(SD_d)^2$).

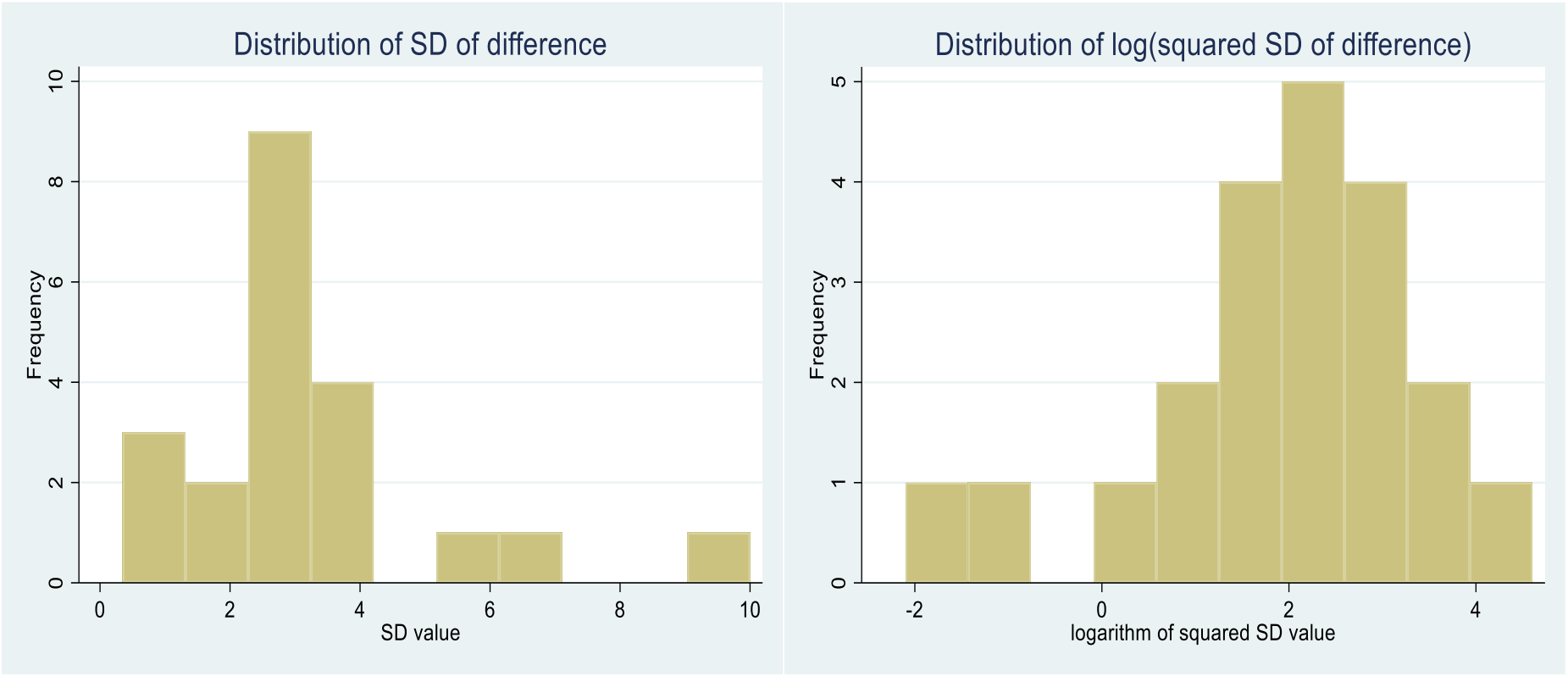
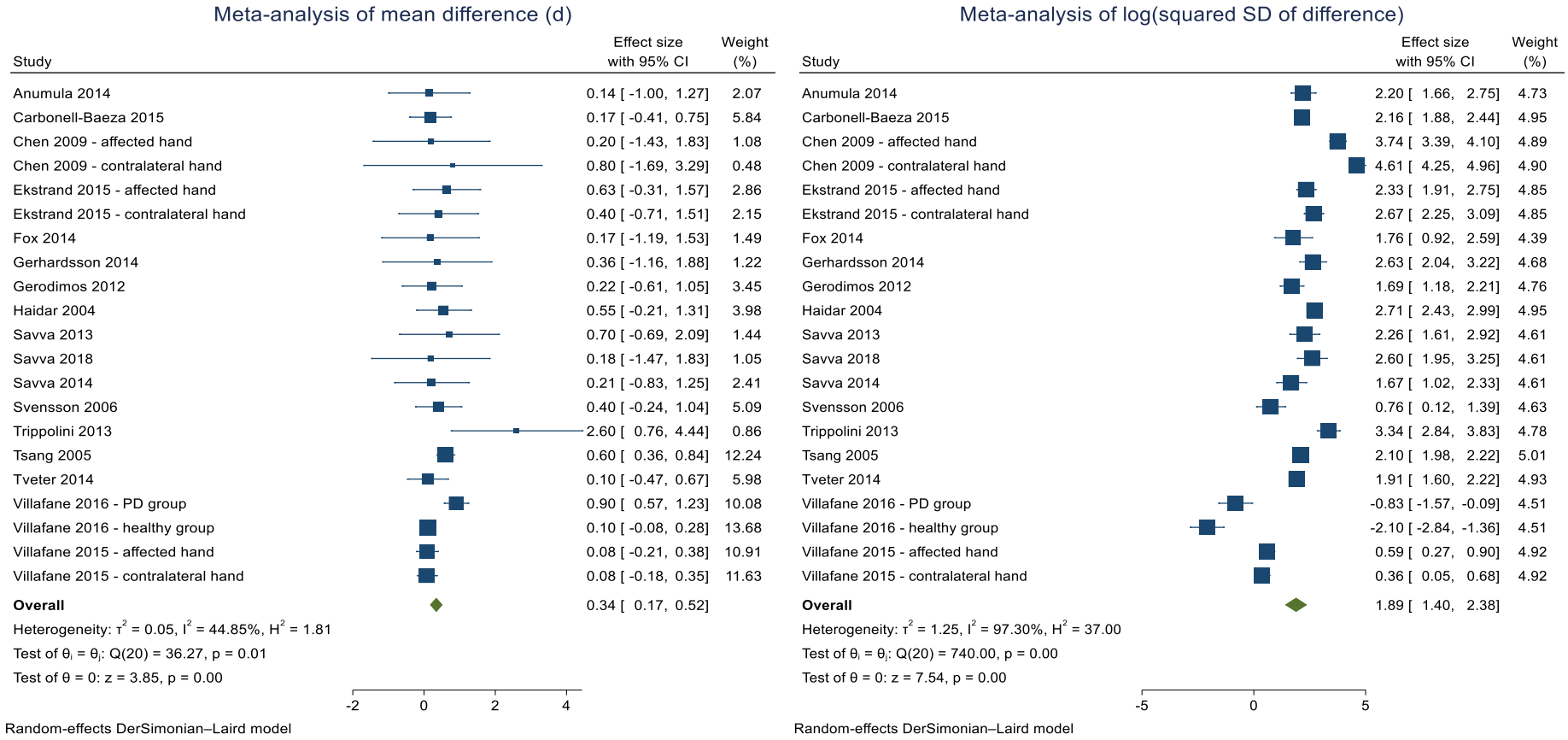


Figure 6.9. Forest plot of the two parameters required for the limits of agreement (d and $\log(SD_d)^2$) including studies examining intra-observer reproducibility across two different testing sessions.

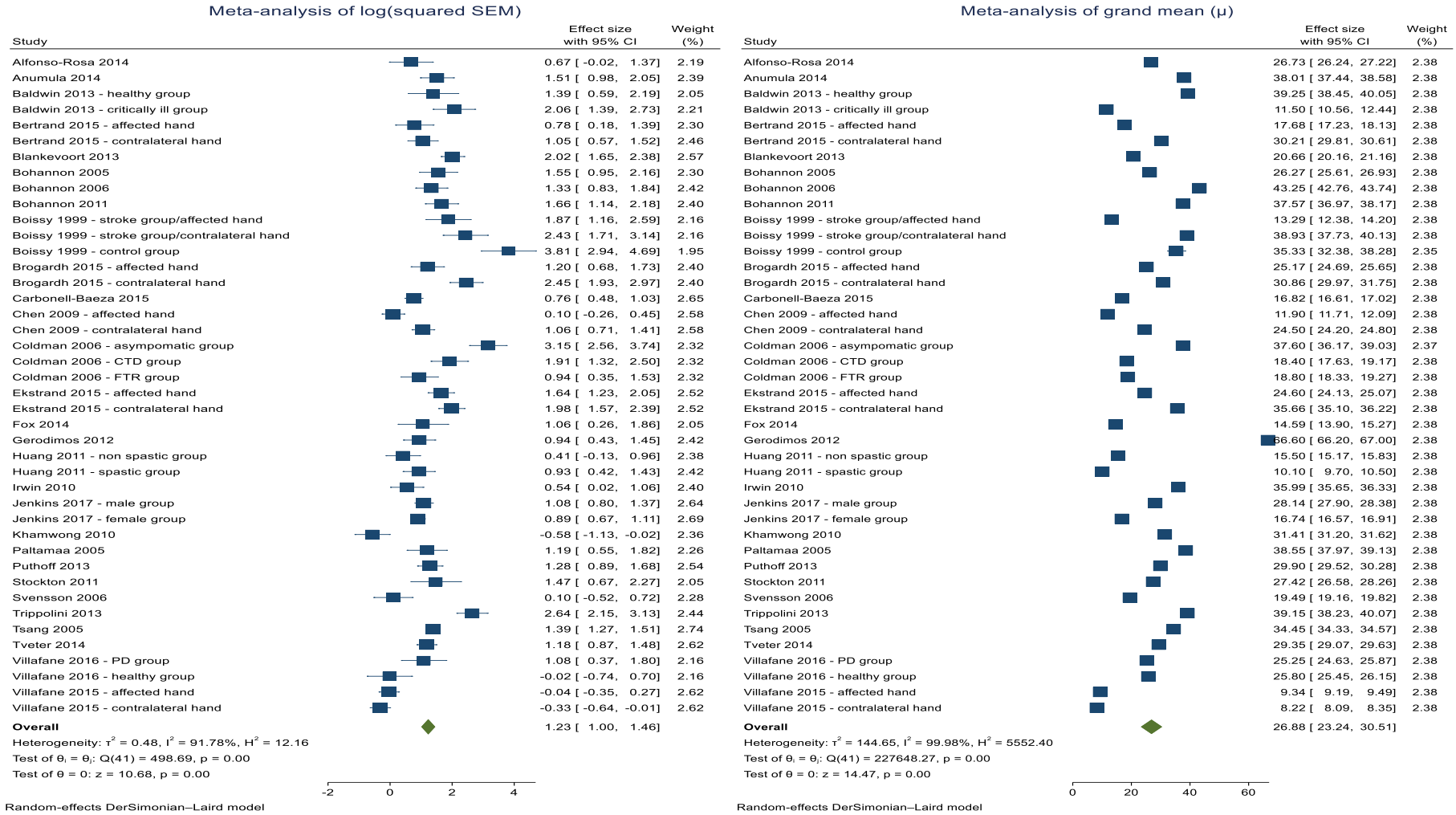


Coefficient of variation

The two parameters required for obtaining the within-patient coefficient of variation were reported, or could be derived from 28 of the 62 studies (45.2%), providing 42 observations included in the meta-analysis each parameter. For both parameters, a high heterogeneity between studies was observed. For the standard error of measurement, the lower and upper quartiles for the transformed values were 0.78 and 1.66 [min= -0.58; max= 3.81]. For the grand mean of the measurements, the lower and upper quartiles for the values reported within studies were 17.7 and 35.7 [min= 8.2; max= 66.6].

The pooled standard error of measurement and grand mean of the measurements across the identified studies were 1.85 [95% CI: (1.65, 2.07)] and 26.88 [95% CI: (23.24, 30.52)], respectively. Using formula 6.20, the corresponding coefficient of variation was 6.90% [95% CI: (2.10%, 9.37%)].

Figure 6.10. Forest plot of the two parameters required for the coefficient of variation (log(SEM^2) and μ) including studies examining intra-observer reproducibility across two different testing sessions.



Other reported statistics

Three studies [188-190] examined the reproducibility of the grip strength measurements using the Spearman correlation coefficient, with the reported estimates ranging from 0.77 to 0.97, while one study [191] reported Kendall's W [192] (i.e., non-parametric test for correlation, alternative to Spearman's) estimates of 0.95 and 0.97 for the ABI group and the healthy group of participants, respectively. Finally, one additional study [193] used the t-test and the linear regression based intercept and slope to examine the systematic differences between the measurements, in addition to other statistics used to examine the random variation in the measurements (such as the Pearson correlation coefficient). For both groups of healthy women and women with non-specific regional pain, the average difference between test and retest was negligible and non-statistically significant, while the linear regression-based intercept and slope were approximately equal to zero and unity, respectively.

Table 6.2. Primary analysis including studies examining the intra-observer reproducibility across 2 testing sessions.

Parameter	**Pooled estimate [95% CI, N observations (participants)]**
Intra class correlation	0.96 [(0.95, 0.97), 68 (3417)]
Pearson correlation	0.96 [(0.94, 0.97), 7 (346)]
Standard error of measurement	1.78 [95% CI: (1.59, 2.00), 54 (2443)]
Smallest detectable change	4.93 [95% CI: (4.40, 5.54), 54 (2443)]
Limits of agreement	0.34 ± 1.96×2.57 [(-4.69, 5.38), 21 (1425)]
Coefficient of variation (%)	6.90% [95% CI: (2.10%, 9.37%)] [42 (2088)]

6.6.3.2. Measurements performed from different observers

Only sixteen studies (20%) examined the inter-observer reproducibility of the grip strength measurements across multiple testing sessions. The results are presented in Table 6.3.

Intra class correlation

The intra class correlation was reported in all 16 studies. Of the 16, 15 studies examined the reliability of grip strength measurements made at two different sessions, with 18 observations included in the meta-analysis [median (Q1, Q3) = 0.97 (0.96, 0.98); min= 0.93; max= 0.99]. The produced weighted average Z value was 2.08 [95% CI: (1.98, 2.18); 95% PI: (1.81, 2.35)]. Using formula 6.7, the corresponding weighted average for the intra class correlation was 0.97 [95% CI: (0.96, 0.98); 95% PI: (0.95, 0.99)].

One additional [194] study examined the reliability of grip strength measurements over 3 different testing sessions, with the intra class correlation being equal to 0.98.

Standard error of measurement and smallest detectable change

The standard error of measurement was reported in 7/16 studies (43.8%), all examining the reproducibility across two testing sessions, with 9 observations included in the meta-analysis [median (Q1, Q3) = 1.70 (1.29, 1.82); min= 0.29; max= 3.37]. The weighted average produced for the logarithm of the squared SEM values was 0.66 [95% CI: (-0.30, 1.63); 95% PI: (-3.27, 4.60)]. The corresponding weighted average values for the standard error of measurement and smallest detectable change were 1.39 [95% CI: (0.86, 2.26); 95% PI: (0.20, 9.98)] and 3.86 [95% CI: (2.38, 6.26); 95% PI: (0.55, 27.64)], respectively.

Limits of agreement

The limits of agreement were reported in 2/16 (12.5%) studies, with 1 observation per study included in the meta-analysis for each of the two parameters required for the calculation of the two limits. The weighted average for the mean difference between the two testing sessions ($\bar{d}$) and the standard deviation of the difference ($SD_{\bar{d}}$) was 0.67 [95% CI: (0.19, 1.15)] and 2.24 [95% CI: (1.75, 2.87)], respectively. The corresponding limits of agreement were 0.67 ± 1.96×2.24.

Coefficient of variation

The two parameters required for obtaining the within-patient coefficient of variation were reported, or possible to be calculated in 3 studies, with 1 observation per study included in the meta-analysis for each parameter. The pooled grand mean of the measurements across the identified studies (μ) and the standard deviation of the measurements made within the patients (SEM) was 25.97 [95% CI: (14.94, 37.03)] and 0.88 [95% CI: (0.24, 3.30)], respectively. Using formula 6.19, the corresponding coefficient of variation was 3.41%.

Table 6.3. Primary analysis including studies examining the inter-observer reproducibility across 2 testing sessions.

Parameter	**Pooled estimate [95% CI, N observations (participants)]**
Intra class correlation	0.97 [0.96, 0.98, 18 (704)]
Pearson correlation	-
Standard error of measurement	1.39 [(0.86, 2.26), 9 (371)]
Smallest detectable change	3.86 [(2.38, 6.26), 9 (371)]
Limits of agreement[1]	0.67 ± 1.96×2.24 [(-), 2 (80)]
Coefficient of variation[1]	3.41% [(-), 3 (159)]

[1]confidence intervals not calculated due to the limited number of observations included in the analysis

6.6.4. Reproducibility of measurements taken within the same testing session

Twelve studies (15%) examined the reproducibility of grip strength measurements made within the same testing session. The results are presented in Table 6.4.

Intra class correlation

The intra class correlation was reported in all 12 studies. Of the 12, 10 studies examined the reproducibility of three consecutive grip strength measurements, with 13 observations included in the meta-analysis [median (Q1, Q3) = 0.96 (0.95, 0.98); min= 0.90; max= 0.99]. The produced weighted average Z-value was 2.03 [95% CI: (1.85, 2.22); 95% PI: (1.33, 2.73)]. The corresponding weighted average ICC estimate was 0.97 [95% CI: (0.95, 0.98); 95% PI: (0.90, 0.99)].

Two additional studies examined the reproducibility of two consecutive grip strength measurements, with 6 observations included in the meta-analysis [median (Q1, Q3) = 0.96 (0.95, 0.96); min= 0.94; max= 0.99]. The produced weighted average Z and ICC values were 1.92 [95% CI: (1.80, 2.03); 95% PI: (1.53, 2.30)] and 0.96 [95% CI: (0.95, 0.97); 95% PI: (0.91, 0.99)], respectively.

Standard error of measurement

The standard error of measurement was reported in 4/12 studies (25%). Of the 4, 3 studies examined the reproducibility of three consecutive grip strength measurements, with 1 observation per study included in the meta-analysis. The remaining study [195] examined the reproducibility of two grip strength measurements in five different cohorts of patients, with the SEM for each cohort approximated through the reported limits of agreement ($SEM = \frac{SD_{\bar{d}}}{\sqrt{2}}$) [15]. The weighted average

SEM value for two and three consecutive within-session measurements were 1.99 [95% CI: (1.74, 2.28)] and 2.68 [95% CI: (2.31, 3.11)], respectively.

Limits of agreement

One study reported the limits of agreement for five different cohorts of patients [195], with 5 observations included in the meta-analysis for each of the two parameters required for the construction of the interval. The weighted average for the mean difference between two consecutive measurements (d) and the standard deviation of the difference (SD_d) was 0.33 [95% CI: (0.16, 0.50)] and 2.82 [95% CI: (2.47, 3.23)], respectively. The corresponding limits of agreement were 0.33 ± 1.96×2.82 = (-5.20, 5.86).

Coefficient of variation

One study [196] examined the reproducibility of two consecutive grip strength measurements produced by the Jamar dynamometer, with the reported coefficient of variation equal to 2.93%.

Table 6.4. Primary analysis of studies examining the reproducibility within testing sessions.

	Within-session reproducibility (2 session measurements)	**Within-session reproducibility (3 session measurements)**
Intra class correlation (Fisher's Z) [95% CI, N observations (participants)]	0.96 [(0.95, 0.97), 6 (1103)]	0.97 [(0.95, 0.98), 13 (923)]
Standard error of measurement [95% CI, N studies (participants)]	1.99 [(1.74, 2.28), 5 (939)]	2.68 [(2.31, 3.11), 3 (114)]

Limits of agreement **[95% CI, N studies (participants)]**	0.33 ± 1.96×2.82 [(-11.38, 12.04), 5 (939)]	-

6.6.5. Subgroup analysis

As described in section 6.5.1, 62/80 studies (77.5%) examined the intra-observer reproducibility of the grip strength measurements. Of the 62, the majority (58/62, 93.5%) examined the reproducibility between two testing sessions, while the remainder used a different number of sessions. Only 16/80 studies (20%) examined the inter-observer reproducibility of the grip strength measurements, while only 12/80 (15%) studies examined the reproducibility of the grip strength measurements made within the same session.

Thus, the analysis was limited to studies examining the intra-observer reproducibility of grip strength measurements between two testing sessions, due to limited data available for inter-observer and within-session reproducibility. Furthermore, the analysis was limited to the intra class correlation and the standard error of measurement only, which were available in 50/58 (86.2%) and 39/58 (67.2%) studies respectively, due to the limited number of estimates reported for any other parameter of reproducibility. The results obtained from the subgroup analysis are presented in Table 6.5.

Summary measures per session

Using the mean of multiple measurements appeared to be the most reliable approach (ICC=0.97 for both two and three measurements) and produced the smallest measurement error, with the mean of three measurements being the most popular approach. Similar values were observed for the

remaining approaches, with the best of three session measurements yielding the largest measurement error (SEM=2.22).

Posture during the assessment

The majority of studies examined a seated position (44, 55%), while patients were tested standing only in 4 (5%) studies. A seated position during the assessment appeared slightly more reliable (ICC=0.97) and produced a lower measurement error (SEM=1.58) compared to a standing position (ICC=0.95, SEM=1.93).

Tested hand in patients with hand injuries

Thirteen studies (16%) examined the reproducibility of the affected (or more affected) hand, with 8 of the 13 studies additionally reporting estimates for the contralateral hand. The produced ICC and SEM values were similar for both hands.

Table 6.5. Subgroup analysis of studies examining the reproducibility across two testing sessions.

Subgroups	**Intra class correlation [(95% CI), N observations (participants)]**	**Standard error of measurement [(95% CI), N observations (participants)]**
Summary measure per session		
Single measurement	0.95 [(0.93, 0.96), 13 (591)]	1.96 [(1.67, 2.31), 11 (274)]
Mean of 2 measurements	0.97 [(0.95, 0.98), 9 (382)]	1.52 [(1.29, 1.80), 8 (343)]
Best of 2 measurements	0.95 [(0.90, 0.98), 7 (408)]	1.98 [(1.53, 2.55), 6 (170)]
Mean of 3 measurements	0.97 [(0.96, 0.97), 41 (2100)]	1.58 [(1.33, 1.87), 31 (1696)]

Best of 3 measurements	0.95 [(0.94, 0.96), 24 (1252)]	2.22 [(1.99, 2.47), 21 (1142)]
Posture during assessment		
Sitting	0.97 [(0.96, 0.97), 55 (2367)]	1.58 [(1.29, 1.92), 44 (3484)]
Standing	0.95 [(0.90, 0.98), 7 (627)]	1.93 [(1.40, 2.67), 4 (112)]
Tested hand in populations with hand injuries		
Affected	0.96 [(0.95, 0.97), 13 (413)]	1.63 [(1.23, 2.16), 13 (413)]
Contralateral	0.96 [(0.94, 0.97), 8 (312)]	1.82 [(1.10, 3.01), 8 (312)]

6.7. Discussion

This systematic review summarises the evidence regarding the reproducibility of grip strength measurements produced by hand-held dynamometers. Eighty primary studies were identified through electronic searching of MEDLINE and EMBASE. The quality of the identified studies was assessed using a recently developed risk of bias tool for studies examining the reliability of clinical outcome measures [37]. In general, the studies were considered to be of good quality. The majority of studies (49/80, 61%) provided doubtful or inadequate evidence that the professionals performing the assessment were blinded to other repeated measurements in the same patient. However, blinding is particularly important when there is subjectivity in the assessment, and thus, less concerning in this case due to the objective nature of the test [197]. For the time interval between measurements made within the same testing session, 26/80 studies (33%) provided insufficient evidence of an appropriate time interval, which may have caused variations in the within-session measurements due to carry-over effects [198]. For the remaining five criteria the risk of bias appeared to be low, with the ratings being at least adequate in ≥90% of the studies.

The majority of the identified studies (62/80, 77.5%) examined the intra-observer reproducibility of grip strength measurements made across two different testing sessions. The most commonly

reported parameter was the intra class correlation (n=54), followed by the standard error of measurement (n=39), the coefficient of variation (n=26), the limits of agreement (n=17), and the Pearson correlation coefficient (n=6). A high level of heterogeneity between studies was noted in the majority of meta-analyses performed (see forest plots under section 6.6.3). This between-study heterogeneity was to some extent expected, given that the overarching aim was to examine the reliability and error of grip strength measurements in general (as opposed to within specific populations), and did not affect the decision to conduct meta-analyses.

The produced weighted average for the intra class correlation was 0.96 (95% CI 0.95, 0.97), suggesting that on average the test is highly reliable for distinguishing patients from each other, i.e., 96.0% of the total variability in the grip strength measurements can be attributed to differences between individuals with only 4.0% attributed to measurement error. The prediction interval for the intra class correlation spanned from 0.80 to 0.99, indicating good to excellent reliability in 95% of all populations potentially considered in the meta-analysis. A very similar estimate was obtained from the meta-analysis of Pearson correlation coefficients, again indicating that the grip strength measurements were produced with high reliability.

The weighted average for the standard error of measurement (SEM) was 1.78 (95% CI: 1.59, 2.00). Assuming normality of the measurements made within patients and homoscedasticity across patients, a repeated measurement is therefore expected on average to lie within 1.78kg of the true grip strength value of the patient. In turn, the estimate for the smallest detectable change (SDC) was 4.93 (4.40, 5.54). If the difference between two within-individual measurements made at two different testing sessions exceeds 4.93kg, it can be stated with 95% confidence that this difference reflects a true change in performance, rather than one anticipated due to measurement error. The 95% prediction intervals for SEM spanned from 0.77kgs to 4.14kgs and from 2.13kgs to 11.47kgs, respectively, suggesting that the measurement error is expected to vary across populations with different characteristics.

The pooled limits of agreement (LoA) ranged from -4.69 to 5.38. If the difference between two measurements produced from the same individual at two different testing sessions falls within this range, it can be stated with 95% confidence that this difference is due to measurement error and not due to a change in performance. The weighted average for the coefficient of variation (CV) was 6.90% (95% CI: 2.10%, 9.37%), which is significantly lower compared to the values considered acceptable in the literature (between 10% and 20%) [108, 129], suggesting low variability in the measurements produced within individuals.

Similar results were obtained when the inter-observer reproducibility across two different sessions and the reproducibility within the same session were examined. The presence of different testers across different measurements sessions did not appear to induce any additional variability, as expected when an objective test is used. For the within-session reproducibility, the measurements were again highly reliable, but the measurement error was larger compared to studies examining the reproducibility between two different testing sessions. However, the results should be interpreted with caution due to the limited number of studies examining the reproducibility of measurements produced within the same testing session.

For the intra-observer reproducibility across two different sessions, the observed between-study heterogeneity did not appear to be attributed to different session summary measures being used, different testing postures of the participants (sitting/standing), or differences between the affected and the contralateral hand being tested in populations with hand injuries. Similar results were observed across the different summary measures and testing postures, with the mean of multiple consecutive grip strength measurements and a sitting position yielding a slightly higher reliability and lower measurement error, compared to any other session summary measure (single session measurement or best of multiple measurements) and a standing position, respectively. These findings are in agreement with the guidelines provided by the American Society of Hand Therapists [123]. Furthermore, for both the affected and contralateral hands, the estimates of reliability were

high and very similar to each other, with a slightly lower measurement error noted for the contralateral hand. It is essential that the measurements produced from each hand are highly reliable, so that robust conclusions can be drawn when the two hands are compared for estimating loss of strength, as the American Society of Hand Therapists recommends [123]. Again, the results obtained from the subgroup analyses should be treated with caution due to the limited number of studies testing populations with hand injuries, or using approaches alternative to those recommended by the American Society of Hand Therapists (e.g., using the maximum of two consecutive measurements as a session summary measure or testing the participants in a standing position).

Strengths and limitations

This systematic review updates the evidence provided from the previous review published in 2011 [124], drawing similar conclusions regarding the reliability of grip strength measurements produced from hand-held dynamometers. The systematic review was conducted using robust review methods. Primary studies were identified by searching two electronic databases (MEDLINE and EMBASE) as current guidelines for conducting systematic reviews recommend [162, 163]. The full text and quality assessment was undertaken independently by two reviewers, while the data extractions were checked by a second reviewer, with any queries being discussed and resolved by consensus. Such methods ensure that potential bias in the review process is minimised, which in turn allows decisions on the reliability and measurement error to be made based on high quality evidence. In addition, compared to the previous review, this review provided quantitative summary evidence for the reliability and error of grip strength measurements. The meta-analytic approaches used for this purpose were very effective in normalizing the distribution of the included study-level estimates. This can be considered a strength, as less biased average estimates are obtained when the distribution of the reported study level estimates is properly modelled. The methods indicated a

high between-study heterogeneity in the majority of meta-analyses performed. Not identifying any factors causing this heterogeneity, and particularly not assessing the effect of the different patient populations included, can be considered a limitation of this study.

6.8. Conclusion

The findings of this review suggest that grip strength measurements obtained from hand-held dynamometers are on average highly reliable and produced with low error. The reliability and measurement error remained high even when methods alternative to the those recommended by the American Society of Hand Therapists were used. However, the between-study heterogeneity observed in the majority of meta-analyses performed suggest that the reliability and measurement error of dynamometers may be reduced in some specific subgroups. In order to provide recommendations for use in practice, further work is required for the identification of any such subgroups.

7. Alternative statistical methods for estimating sources of variability in the measurements of count-based biomarkers

7.1. Introduction

So far, the thesis has been concerned with statistical methods for estimating different sources of variability of biomarkers, measured in a continuous scale. As described in Chapter 2, these methods assume underlying normality of the measurements for each potential level of variability. However, biomarkers may often be expressed as counts (i.e., whole numbers) or rates (i.e., counts measured in relation to another quantity) rather than continuous measurements. Examples include the number of foci per labial salivary glandular area for evaluating Sjogren's syndrome [199], or the number of white blood cells per litre for evaluating the severity of COPD [200]. In this case, the distribution of the produced measurements is likely to be highly skewed [201], which means the methods described in Chapter 1 are no longer applicable due to the violation of the normality assumption. Many researchers suggest the use of the log-transformation as a potential solution to this problem [27, 56]. However, evidence suggests that this transformation does not always lead to a better approximation of normality [106], while the transformation cannot be applied when count-measurements take the value of zero.

7.2. Aim

To present alternative statistical methods for estimating different sources of variability in measurements produced between and within-individuals, which are potentially more appropriate for count-based biomarkers.

7.3. Statistical models used for analysing multi-level count data

This section reviews the two most widely applied statistical distributions for modelling count-based data. These include the Poisson distribution and the negative binomial distribution [202].

7.3.1. The Poisson distribution

The most popular distribution used for the analysis of count data is the Poisson distribution. As the name implies, this distribution was first described by Denis Poisson, in 1837 [203]. When used for data with multiple hierarchical levels of variability, a key assumption that this distribution holds is that the lower-level variance equals the mean of the produced measurements [202].

7.3.1.1. The two-level Poisson regression model

If the equality assumption is met, the Poisson distribution can be used to model the probability of the y_{ij} response occurring with the function

$$P(Y = y_{ij}) = \frac{e^{-y_{ij}}(\lambda_{ij})^{y_{ij}}}{y_{ij}!}, \tag{7.1}$$

where y_{ij} and λ_{ij} denote the observed and expected count for the i_{th} measurement $(i = 1, \dots, n_i)$ within the j_{th} individual $(j = 1, \dots, n_j)$, respectively. The corresponding model for estimating λ_{ij} is mathematically expressed as

$$y_{ij}|\lambda_{ij} \sim Poisson(\lambda_{ij}) \tag{7.2}$$

$$\log(\lambda_{ij}) = \beta_0 + \log(exposure_{ij}) + u_j,$$

where $u_j \sim N(0,\ \sigma_u^2)$ denotes the random effects parameter for the upper-level variability (i.e., the between-individual biological variability), while $\log(exposure_{ij})$ indicates that the modelled response is a rate, meaning the produced counts are measured per unit of an "exposure" variable (e.g., number of nodes per sample area, number of heart beats per minute). If the exposure is the

same for every measurement, the term $\log(exposure_{ij})$ may be omitted from the model. The two parameter estimates obtained from the above regression model are

- the regression intercept β_0 of $\log(y_{ij})$.
- the variance σ_u^2 for the between-individual variability in $\log(y_{ij})$.

Both parameters are estimated on the natural logarithm scale, due to the canonical log-link function that this model uses [202, 204].

Estimation of grand mean and different components of variance

Statistical methods for estimating the mean and different components of variance of the measurements in the original count scale (rather than the logarithmic) are presented in Leckie et al [202] and Austin et al [204]. The expected mean value of y_{ij} (averaged over u_j) is estimated as

$$\mu = E[Y] = E\left[e^{\beta_0 + u_j}\right] \Rightarrow \textit{(given that } e^{\beta_0} \textit{ is constant)}$$
$$\mu = e^{\beta_0} \times E[e^{u_j}]$$

For $E[e^{u_j}]$, Austin et al [204] present a calculation based on the integration formula provided in Spiegel [205]. For any random variable $u \sim N(0, \sigma_u^2)$, it follows that

$$E[e^u] = E\left[\int_{-\infty}^{+\infty} e^u \left(\sqrt{\frac{1}{2\pi}}\, e^{\frac{-u^2}{2}}\right) du\right] = e^{\frac{\sigma_u^2}{2}} \qquad (7.3)$$

Based on equation 7.3, the expected mean of y_{ij} (or the marginal expectation, as termed in Leckie et al [202]) is equal to

$$\mu = e^{\beta_0 + \frac{\sigma_u^2}{2}} \qquad (7.4)$$

For the expected total variance (or marginal variance [202]) of y_{ij}, Leckie et al [202] provide a calculation based on McCulloch and Searle (page 12) [206], as follows

$$Var[Y] = Var\{E[y_{ij}|\beta_0, u_j]\} + E\{Var[y_{ij}|\beta_0, u_j]\} \Rightarrow$$

$$Var[Y] = E\left\{\left(E[y_{ij}|\beta_0, u_j]\right)^2\right\} - \left(E\{E[y_{ij}|\beta_0, u_j]\}\right)^2 + E\{E[y_{ij}|\beta_0, u_j]\} \Rightarrow$$

$$Var[Y] = E\left[e^{2\beta_0 + 2u_j}\right] - \left(E\left[e^{\beta_0 + \frac{u_j}{2}}\right]\right)^2 + E\left[e^{\beta_0 + \frac{u_j}{2}}\right] \Rightarrow$$

$$Var[Y] = e^{2\beta_0 + 2\sigma_u^2} - e^{2\beta_0 + \sigma_u^2} + e^{\beta_0 + \frac{\sigma_u^2}{2}} \Rightarrow$$

$$Var[Y] = \underbrace{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}_{①} + \underbrace{(\mu)}_{②} \tag{7.5}$$

With respect to decomposing the above into the different variance components, the second term is equal to the expected lower-level variance, which corresponds to the variability in the count-measurements produced within individuals. This is based on the assumption of the equality between the mean and lower-level variance that the Poisson distribution holds. The remainder represents the upper-level variance which corresponds to the true variability between the individuals [202, 204].

7.3.1.3. The three-level Poisson regression model

As described in Chapter 2, studies are often interested in examining three hierarchical levels of variability in the test results (or equally, two levels of measurement error). For example, when studies examine the variability of laboratory-based tests, the recruited individuals (upper level) produce multiple samples at different time points (middle level), which are in turn assessed twice (lower level). In such cases, the model described under section 7.3.1.2 (equation 7.2), is extended to

$$y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk}) \tag{7.6}$$

$$\log(\lambda_{ijk}) = \beta_0 + \log(exposure_{ijk}) + u_k + v_j,$$

where $u_k \sim N(0,\ \sigma_u^2)$ is the random effects parameter for the upper-level variability (i.e., true between-individual variability), and $v_j \sim N(0,\ \sigma_v^2)$ is the random effects parameter for the middle-level variability, which corresponds to the upper-level measurement error (e.g., for laboratory tests, the true variability in measurements produced within individuals over time). The three parameter estimates obtained from the above regression model are

- the regression intercept β_0 of $\log(y_{ijk})$.
- the variance σ_u^2 for the upper-level variability in $\log(y_{ijk})$.
- the variance σ_v^2 for the middle-level variability in $\log(y_{ijk})$.

Estimation of grand mean and different components of variance

When a three-level Poisson model is used, the calculation of the expected mean and corresponding total variance of y_{ijk} (given β_0 and averaged over u_k and v_j) is now extended to

$$\mu = E[Y] = e^{\left(\beta_0 + \frac{\sigma_u^2}{2} + \frac{\sigma_v^2}{2}\right)}, \tag{7.7}$$

and

$$Var[Y] = \underbrace{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}_{①} + \underbrace{(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right)}_{②} + \underbrace{(\mu)}_{③} \tag{7.8}$$

where ① represents the upper-level variance, ② represents the middle-level variance, and ③ represents the lower-level variance [202].

7.3.2. The negative binomial distribution

As mentioned under section 7.3.1, the use of the Poisson distribution for modelling multilevel data requires the lower-level variance to equal the mean value of the produced measurements. However,

it has been empirically shown that this assumption is often violated in practice, with the observed lower-level variance being larger than that expected by the Poisson distribution. This phenomenon is called *overdispersion* [202]. An alternative distribution for modelling count data, able to additionally accommodate any potential overdispersion observed in the data, is the negative binomial distribution. Student was the first to propose this distribution for counting red blood cells using the hemocytometric camera [207]. This method consists of stretching blood drops on a special slide which has a camera with a grid superimposed. The specimen is then read on a microscope by a reader, who counts the number of red blood cells in a fixed number of quadrats. In this case, excess variability was caused due to the spread of the blood on the slide being more heterogeneous than what the Poisson distribution would expect.

7.3.2.1. The two-level negative binomial regression model

For data with two potential levels of variability (i.e., individuals and measurements produced within each individual), a negative binomial model can be expressed as

$$y_{ij}|\lambda_{ij} \sim Poisson(\lambda_{ij}) \tag{7.9}$$

$$\log(\lambda_{ij}) = \beta_0 + \log(exposure_{ij}) + u_j + \omega_{ij}$$

Compared to the Poisson model, this model includes an additional parameter ω_{ij} for capturing any excess lower-level variability that is beyond the Poisson process, with $e^{\omega_{ij}} \sim Gamma\left(\frac{1}{a}, a\right)$ distributed around a mean of 1 with variance a. The three parameter estimates obtained from the above regression model are

- the regression intercept β_0 of $\log(y_{ij})$.
- the variance σ_u^2 for the upper-level variability in $\log(y_{ij})$.
- the variance a for any excess lower-level variability in y_{ijk}, with higher values indicating a greater overdispersion. If $a = 0$, this model is equivalent to a Poisson model.

All parameters are again estimated on the logarithm scale, as this model also uses the canonical log-link function [202].

Estimation of grand mean and different components of variance

The calculation of the expected mean value of y_{ij} (averaged over u_j and ω_{ij}) is identical to that for the two-level Poisson model (see equation 7.4). However, the corresponding lower-level variance is no longer equal to the mean, and is instead calculated as

$$Var[y_{ij}|u_j] = \mu(1 + a \times \mu), \tag{7.10}$$

where a is the estimated overdispersion parameter, and μ is calculated as above. The variance for the upper level is calculated in the same way as for the two-level Poisson model (see ① in equation 7.5) [202].

7.3.2.2. The three-level negative binomial regression model

When modelling data with three potential levels of variability, the model extends to

$$y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk}) \tag{7.11}$$
$$\log(\lambda_{ijk}) = \beta_0 + \log(exposure_{ijk}) + u_k + v_j + \omega_{ijk},$$

where $u_k \sim N(0,\ \sigma_u^2)$ and $v_j \sim N(0,\ \sigma_v^2)$ are the random effects parameters for the upper and middle levels of variability, respectively, and ω_{ijk} is the overdispersion parameter, defined as above. The four parameter estimates obtained from the above regression model are

- the regression intercept β_0 of $\log(y_{ijk})$.
- the variance σ_u^2 for the upper-level variability of $\log(y_{ijk})$.
- the variance σ_v^2 for the middle-level variability of $\log(y_{ijk})$.

- The variance a for any excess lower-level variability in y_{ijk}.

Estimation of grand mean and different components of variance

The calculation of the expected mean value of y_{ijk} (averaged over u_k, v_j, and ω_{ijk}) is identical to that for the three-level Poisson model (see equation 7.7). The lower-level variance is again calculated using equation 7.10, while the variance for the remaining levels is calculated in the same way as for the three-level Poisson model (see ①, ② in equation 7.8) [202].

7.3.3. Which model should be used?

So far it has been shown that a Poisson model is nested within a negative binomial model, as a negative binomial model includes an additional parameter which accounts for potential overdispersion. Therefore, the decision on which model should be chosen is entirely dependent on the presence of significant overdispersion in the data. If present, the negative binomial model should be preferred, while if not present, the Poisson model may be employed.

The likelihood ratio test

A formal statistical investigation of whether significant overdispersion is present, is possible using the likelihood ratio test [202]. This is a frequently used test for comparing the change in deviances of two nested models (for each model, this is given by minus twice the maximum value of the log likelihood function) [208]. Assuming that we have two models, one with p parameters nested within an alternative with $p + 1$ parameters, then the models may be compared by testing weather the additional parameter is significantly different from zero. Therefore, under the null hypothesis that it is not, the following hypothesis test may be used

$$-2\log\left(\frac{L(p)}{L(p+1)}\right) \sim \chi_1^2, \tag{7.12}$$

where $L(p)$ and $L(p+1)$ is the maximum values of the log likelihood function produced from a model with p and $p+1$ parameters, respectively, and χ_1^2 represents a chi-square distribution with one degree of freedom.

7.4. Parameters of reliability and measurement error

This section presents four parameters of reliability and measurement error of count-based biomarkers, which can be calculated using the estimates produced from the different models described under section 7.3. Known parameters, which are also used for evaluating the reliability and measurement error of continuous measurements, include the standard error of measurement, the intra class correlation, and the coefficient of variation. A new parameter, potentially useful for evaluating the reliability of count-based measurements, which has not been introduced so far in the thesis, includes the median rate ratio. All parameters are presented in Table 7.1.

Table 7.1. Estimation of parameters for each statistical model presented.

Parameter	Poisson model	Negative binomial model
- *Two-level count response model*		
Grand mean (μ)	$e^{\left(\beta_0 + \frac{\sigma_u^2}{2}\right)}$	$e^{\left(\beta_0 + \frac{\sigma_u^2}{2}\right)}$
Lower-level variance	μ	$\mu(1 + a \times \mu)$
Upper-level variance	$(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)$	$(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)$
Standard error of measurement (SEM)	$\sqrt{\mu}$	$\sqrt{\mu(1 + a \times \mu)}$
Coefficient of variation (CV)	$(\mu)^{-\frac{1}{2}} \times 100$	$\frac{\mu(1 + a \times \mu)}{\mu} \times 100$
Intra class correlation (ICC)	$\frac{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right) + \mu}$	$\frac{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right) + \mu(1 + a \times \mu)}$
Median rate ratio (MRR)	$e^{\left(\sqrt{2\sigma_u^2}\ \Phi^{-1}(0.75)\right)}$	–
- *Three-level count response model*		
Grand mean (μ)	$e^{\left(\beta_0 + \frac{\sigma_u^2}{2} + \frac{\sigma_v^2}{2}\right)}$	$e^{\left(\beta_0 + \frac{\sigma_u^2}{2} + \frac{\sigma_v^2}{2}\right)}$
Lower-level variance	μ	$\mu(1 + a \times \mu)$
Middle-level variance	$(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right)$	$(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right)$
Upper-level variance	$(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)$	$(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)$
Standard error of measurement (SEM)	$\sqrt{(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right) + \mu}$	$\sqrt{(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right) + \mu(1 + a \times \mu)}$
Coefficient of variation (CV)	$\frac{\sqrt{(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right) + \mu}}{\mu} \times 100$	$\frac{\sqrt{(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right) + \mu(1 + a \times \mu)}}{\mu} \times 100$
Intra class correlation (ICC)	$\frac{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right) + (\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right) + \mu}$	$\frac{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right) + (\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right) + \mu(1 + a \times \mu)}$

β_0 is the regression intercept obtained from each model, σ_u^2 is the upper-level variance obtained from each model,

σ_v^2 is the middle-level variance obtained from a 3-level model, a is the excess lower-level variance obtained from a negative binomial model.

7.4.1. The standard error of measurement (SEM)

As described in Chapter 1, the standard error of measurement quantifies the variability of multiple measurements produced within individuals. However, the calculation of the parameter may change, depending on the statistical model being used for estimating the variability in the count-based measurements.

SEM based on Poisson model

When using a two-level model (i.e., only one source of measurement error is considered), the standard error of measurement is equal to the standard deviation of multiple measurements produced within-individuals, calculated as

$$SEM = \sqrt{e^{\beta_0 + \frac{\sigma_u^2}{2}}} = \sqrt{\mu} \tag{7.13}$$

When two sources of within-individual variability are formally examined using a three-level model the calculation extends to the square root of the sum of variances attributed to each source. In this case, this is equal to

$$SEM = \sqrt{\underbrace{(\mu)^2(e^{(\sigma_u^2)})\left(e^{(\sigma_v^2)} - 1\right)}_{①} + \underbrace{(\mu)}_{②}}, \tag{7.14}$$

where ① is the upper level of variability within individuals (e.g., for laboratory tests, the within-individual biological variability), and ② is the lower level of variability within individuals (e.g., for laboratory tests, the analytical variability).

SEM based on negative binomial model

When a negative binomial model is used, the standard error of measurement is calculated in a similar fashion to a Poisson model. However, as described in section 7.3.2, the calculation of the

lower-level variance alters so that it accounts for any potential overdispersion being present. Thus, for a two-level negative binomial model, the calculation of the SEM is extended to

$$SEM = \sqrt{\mu(1 + a \times \mu)}, \tag{7.15}$$

while for a three-level negative binomial model, the calculation is extended to

$$SEM = \sqrt{\underbrace{(\mu)^2(e^{(\sigma_u^2)})(e^{(\sigma_v^2)} - 1)}_{①} + \underbrace{\mu(1 + a \times \mu)}_{②}}, \tag{7.16}$$

where ① is the upper level of variability within individuals (e.g., for laboratory tests, the within-individual biological variability), and ② is the lower level of variability within individuals (e.g., for laboratory tests, the analytical variability).

7.4.2. The coefficient of variation (CV)

As descried in Chapter 1, the coefficient of variation allows the total variability within individuals to be expressed in relation to the grand mean of the measurements. For count-based measurements, the calculation is as follows

$$CV = \frac{SEM}{\mu} \times 100, \tag{7.17}$$

where SEM is calculated as describe in section 7.4.1, and μ is estimated using equation 7.4 (if a two-level model is used) or equation 7.7 (if a three-level model is used).

7.4.3. The intra class correlation (ICC)

The intra class correlation expresses the reliability of the test measurements, and is calculated as the proportion of the total variance attributed to true differences between individuals [202, 204]. For count-based measurements, this is mathematically expressed as

$$ICC = \frac{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right)}{(\mu)^2\left(e^{(\sigma_u^2)} - 1\right) + SEM^2}, \tag{7.18}$$

where the numerator (and first term of the denominator) represents the true variance between individuals, and SEM is again calculated as describe in section 7.4.1.

7.4.4. The median rate ratio (MRR)

This chapter introduces the median rate ratio as a potentially useful parameter for expressing the reliability of a test producing count-based measurements (i.e., the ability of a test to distinguish individuals with a better outcome to those with a worse outcome, despite the presence of measurement error). In contrast to the aforementioned parameters, the calculation of this parameter is possible only when a two-level Poisson model is used [204]. The parameter denotes the median relative difference in the rate between a randomly selected count-based measurement produced from an individual with an overall higher rate, and a randomly selected count-based measurement produced from an individual with an overall lower rate [204]. For given values of the regression intercept and the exposure, the expected rate ratio of 2 randomly selected count-based measurements, i and i', from 2 randomly selected individuals, j and j', can be calculated as

$$RR = \frac{e^{\beta_0 + \log\left(exposure_{ij}\right) + u_j}}{e^{\beta_0 + \log\left(exposure_{ij}\right) + u_{j'}}} = e^{|u_j - u_{j'}|} \tag{7.19}$$

If all such pairwise comparisons available in the data are considered, the median value of the produced distribution of rate ratios can then be approximated as

$$MRR = e^{\left(\sqrt{2\sigma_u^2}\ \Phi^{-1}(0.75)\right)}, \tag{7.20}$$

where Φ^{-1} represents the inverse of the standard normal cumulative distribution, and σ_u^2 represents the estimated between-individual variance of the log-rate. This estimate is produced directly from the two-level Poisson model (see section 7.3.1). Larger values produced for the median

rate ratio correspond to higher between-individual variability, while a value of 1 indicates no variability between individuals.

Austin et al [204] state that this concept arises from the fact that the distribution of $|u_j - u_{j'}|$ produced from all available pairwise comparisons is half normal with variance equal to $2\sigma_u^2$, and the median value of this distribution is given by $\sqrt{2\sigma_u^2}\ \Phi^{-1}(0.75)$.

7.5. Discussion

This chapter introduces statistical methods available for estimating sources of variability in the measurements of biomarkers, expressed as counts or rates. The two statistical models available for this purpose include the Poisson random-effects and negative binomial random-effects models. Several parameters of reliability and measurement error can in turn be calculated based on the estimates produced from the two models. Known parameters include the standard error of measurement, the coefficient of variation, and the intra class correlation, while a new, potentially useful parameter of reliability (i.e., the median rate ratio) was also introduced. In contrast to the standard methods available for estimating sources of variability, the use of these methods no longer requires a normal distribution for the measurements produced at different levels. Thus, compared to the standard normality-based methods, these methods are likely to produce less biased estimates of reliability and measurement error of count measurements, where the assumption of normality is known to often be violated. Less biased estimates will in turn allow researchers to make more robust decisions regarding the use of a count-based test in clinical practice. However, calculating the variability of different levels is more computationally intensive compared to the standard methods used for continuous measurements, which can be considered a limitation of these methods. The next chapter illustrates the application and interpretation of these methods using a cohort of patients with confirmed primary Sjogren's syndrome as a case study.

7.6. Conclusion

This chapter presents alternative methods for estimating test reliability and measurement error in primary studies. In contrast to the standard approach based on linear regression, the use of these methods no longer requires the measurements produced at different variability levels to follow a normal distribution. Thus, these methods are likely to provide less biased estimates of variability for count measurements, where the assumption of normality is known to often be violated. Less biased variability estimates will in turn help researchers draw more robust conclusions on the reliability and measurement error of a count-based test, and therefore whether the test is fit for use in clinical practice.

8. Application of statistical methods appropriate for estimating sources of variability in count-based biomarkers

8.1. Introduction

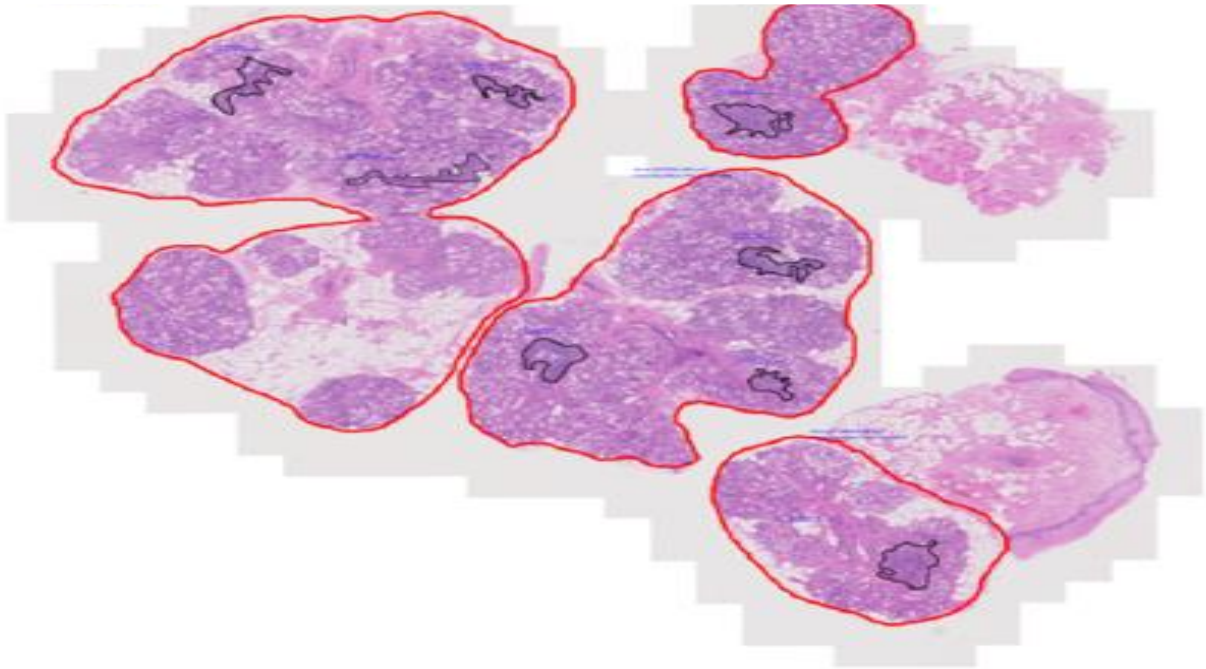
Alternative methods for examining the reliability and measurement error of count-based biomarkers were introduced in Chapter 7. In this chapter, these methods were applied into a data set of 32 patients with confirmed primary Sjogren's syndrome who underwent labial salivary gland biopsy. The data was collected as part of the Optimising Assessments in Sjogren's Syndrome (OASIS) study, with all patients recruited in the study fulfilling the 2016 ACR/EULAR classification criteria [209]. The biomarker of interest was the focus score, which is commonly used for the diagnosis and classification of primary Sjogren's syndrome (PSS) [199, 210], and has also functioned as surrogate endpoint in clinical trials [211].

8.2. Clinical background

Sjogren's syndrome (SS), named after the Swedish ophthalmologist Henrik Sjögren in 1933 [212], is an autoimmune disorder caused by the lymphocytic infiltration of exocrine glands, which results in dysfunction of the salivary and lacrimal glands. The syndrome can be classified into two types, primary and secondary Sjogren's syndrome. Primary Sjogren's syndrome (PSS) occurs in the absence of other autoimmune diseases and is characterised by keratoconjunctiva sicca (dry eyes) and xerostomia (dry mouth), collectively called the sicca syndrome. In contrast, Secondary Sjogren's syndrome (SSS) is presented along with other autoimmune diseases, most commonly with rheumatoid arthritis, but also with polymyositis, polyarteritis nodosa, progressive systemic sclerosis (scleroderma), and systemic lupus erythematosus [213]. The clinical presentation of PSS may vary. The onset is usually at age 40-60 years, with a 9:1 female to male ratio [214]. Salivary gland

histopathology, most commonly performed on labial salivary gland (LSG) biopsies, contributes to the clinical classification and diagnosis of PSS. The most characteristic feature of PSS on LSG biopsy is the focus score, defined as the total number of foci per $4mm^2$ of salivary gland tissue [210]. According to the American College of Rheumatology (ACR)-European League Against Rheumatism (EULAR) a positive histopathology finding based on a focus score≥1 is a requirement for the diagnosis of PSS in the absence of anti-Ro/SSA antibodies [215]. Moreover, evidence suggests that the focus score may also have potential as a biomarker in clinical trials [210].

Figure 8.1. Microphotograph illustrating LSG biopsy obtained from a patient with confirmed disease. The total number of foci (black outlined area) is 8. The measured glandular area (in red) is 20.89mm^2. This gives a focus score of 1.53 for the patient. Graph taken from Fisher et al [210].



8.3. Components of variability in the focus score

The two potential components of variability in the focus score include:

- **Between-patient variability.** That is, any variability in the focus score attributed to true biological differences across the recruited individuals.
- **Between-gland within-patient variability.** That is, any variability in the focus score attributed to differences between the glands observed within each patient's biopsy (e.g., some glands may be larger, or may contain a higher number of foci, compared to others).

8.4. Objectives

To apply statistical methods appropriate for count-based biomarkers in order to estimate potential sources of variability in the focus score, and compare the performance of these methods to the standard approach used for continuous measurements.

8.5. Statistical methods

The baseline characteristics of the 32 patients recruited in the OASIS study were tabulated using appropriate summary statistics. The focus score was calculated for each gland observed within each patient's biopsy, as

$$Focus\ score = \frac{(Number\ of\ foci\ observed\ in\ the\ gland)}{(Area\ of\ the\ gland\ in\ mm^2)} \times 4mm^2 \qquad (8.1)$$

A histogram was used to assess the distribution of the two parameters required for the calculation of the focus score (number of foci within glands and the area of the glands), while a dot plot was used to visually inspect the observed variability in the focus score between the glands, and between the patients. Several multilevel regression models were then employed to estimate the two potential components of variability in the focus score. Initially, the focus score produced for each gland was treated as a rate (i.e., number of foci per $4mm^2$ of glandular area) and analysed using both a Poisson model and a negative binomial model. Subsequently, the focus score of each gland

was analysed as a continuous measurement using a linear regression model. All models included a random effects parameter to estimate any potential variability attributed to true differences between patients. A formal assessment of how well each model fitted the data, and therefore estimated the two components of variability, was carried out using the AIC and BIC criteria (see section 8.5.4). The measurement error and reliability of the focus score were in turn expressed using the standard error of measurement, the coefficient of variation, the intra class correlation, and the median rate ratio.

8.5.1. Analysis using a two-level Poisson model

A two-level Poisson model can be expressed as

$$\log(foci_{ij}) = \beta_o + \log(area_{ij}) + u_j \qquad (8.2)$$

where $foci_{ij}$ denotes the expected number of foci within the i_{th} gland of the j_{th} patient ($j = 1, \dots, 32$), β_o is the regression intercept, $size_{ij}$ is the area of i_{th} gland within the j_{th} patient, and $u_j \sim N(0,\ \sigma_u^2)$ is the patient-level random effects parameter assumed normally distributed around a mean of 0 with variance σ_u^2. The two parameter estimates obtained from the above regression model are:

- the regression intercept β_0 of the logarithm of the foci rate (i.e., number of foci per 1 mm^2 of glandular area).
- the variance σ_u^2 for the upper-level variability in the logarithm of the foci rate.

Estimation of grand mean and variance components

The mean focus score was estimated as

$$\mu = e^{\left(\beta_o + \log 4 + \frac{\sigma_u^2}{2}\right)}, \quad (8.3)$$

where β_o and σ_u^2 are defined as above, while $(\beta_o + \log 4)$ is now the extended regression intercept corresponding to the logarithm of the number of foci per $4mm^2$ of glandular area. By the assumption of equality between the expected mean and lower-level variance, the value produced from equation 8.3 is also equal to the variance expected due to differences in the focus score between the glands within patients $\left(= Var_{glands}\right)$. The expected between-patient variance of the focus score was in turn estimated as

$$Var_{patients} = (\mu)^2\left(e^{(\sigma_u^2)} - 1\right) \quad (8.4)$$

8.5.2. Analysis using a two-level negative binomial model

A two-level negative binomial model is expressed as

$$\log\left(foci_{ij}\right) = \beta_o + \log\left(area_{ij}\right) + u_j + \omega_{ij} \quad (8.5)$$

where ω_{ij} is the additional parameter for capturing any excess lower-level variability that is beyond the Poisson process, with $e^{\omega_{ij}} \sim Gamma(\frac{1}{a}, a)$ distributed around a mean of 1 with variance a. The three parameter estimates obtained from the above regression model are:

- the regression intercept β_0 of the log-foci rate.
- the variance σ_u^2 for the upper-level variability in the log-foci rate.
- the variance a for any excess lower-level variability in the foci rate.

Estimation of grand mean and variance components

The calculation of the mean focus score (μ) and the patient-level variance ($Var_{patients}$) were identical to the Poisson model, using equations 8.3 and 8.4, respectively. However, the calculation of the gland-level variance was this time extended to account for any potential overdispersion, as follows

$$Var_{glands} = \mu(1 + \alpha \times \mu) \quad (8.6)$$

8.5.3. Analysis using a two-level linear model

A two-level linear model can be expressed as

$$Focus\ score_{ij} = \beta_o + u_i + e_{ij}, \quad (8.7)$$

where $Focus\ score_{ij}$ is calculated using equation 8.1, β_o is the regression intercept, $u_i \sim N(0,\ \sigma_u^2)$ is the random effects parameter for the between-patient variability, and $e_{ij} \sim N(0,\ \sigma_e^2)$ is the random error term for the variability between the glands within patients. The three parameter estimates obtained from the above regression model are:

- the regression intercept β_0 of the focus score.
- the variance σ_u^2 for the patient-level variability in the focus score.
- the variance σ_e^2 for the gland-level variability in the focus score.

Estimation of grand mean and variance components

In contrast to the Poisson and the negative binomial models, the grand mean and variance components are this time estimated directly from the model, and expressed in the original scale (rather than the logarithmic). Thus, the estimates of β_0, σ_u^2 and σ_e^2 correspond to the expected

mean (μ), between-patient variance ($Var_{patients}$), and between-gland within-patient variance of the focus score (Var_{glands}), respectively.

Table 8.1. Estimation of the grand mean and variance components across the three models.

	Linear model	Poisson model	Negative binomial model
Grand mean (μ)	β_o	$e^{\beta_o + \log 4 + \frac{\sigma_u^2}{2}}$	$e^{\beta_o + \log 4 + \frac{\sigma_u^2}{2}}$
Gland-level variance	σ_e^2	$e^{\beta_o + \log 4 + \frac{\sigma_u^2}{2}} (= \mu)$	$\mu(1 + \alpha \times \mu)$
Patient-level variance	σ_u^2	$(\mu)^2(e^{(\sigma_u^2)} - 1)$	$(\mu)^2(e^{(\sigma_u^2)} - 1)$

8.5.4. Method for comparing the fitted models

The most popular criteria used for comparing how well different statistical models fit the observed data include the Akaike's (AIC) and the Bayesian information criteria (BIC) [216]. The AIC and BIC values for a model can be calculated as

$$AIC = -2\log(L) + 2 \times q \tag{8.8}$$

and

$$BIC = -2\log(L) + logN \times q \tag{8.9}$$

were L denotes the maximum value of the log likelihood function produced from each model, q is the number of parameters estimated in each model, and N is the total number of observations included in the analysis.

For both information criteria, a lower value indicates a better model performance. Both AIC and BIC values are highly dependent on the deviance value (i.e., first term of equations 8.8 and 8.9) each model yields, as a smaller deviance value indicates a better model performance. Furthermore, in support of parsimony, both criteria penalise for the number of parameters estimated in the model.

Thus, a model with a lower number of estimated parameters would most likely be chosen over a model with a higher number of estimated parameters.

8.5.5. Parameters of reliability and measurement error

Several statistical parameters were then calculated based on the estimates produced from each model, in order to examine the reliability and measurement error of the focus score. Parameters of reliability included the intra class correlation and the median rate ratio, while parameters of measurement error included the standard error of measurement and the coefficient of variation. All parameters were presented along with 95% confidence intervals, which were constructed via multilevel bootstrapping with bias-correction (method introduced in Chapter 2) [79, 80, 125, 126]. For each parameter, a sampling distribution was obtained by fitting the model to 1000 bootstrapped samples [79, 126]. The 2.5% and 97.5% percentiles of the produced sampling distribution were used as the lower and upper confidence bound, respectively, and were adjusted in the appropriate direction in case the original model-based estimate did not lie at the 50^{th} percentile.

Calculation of the standard error of measurement

The standard error of measurement was in this case equal to the standard deviation of the measurements of the glands within each patient. The parameter was calculated for each model as

$$SEM = \sqrt{Var_{glands}} \tag{8.10}$$

Calculation of the coefficient of variation

The coefficient of variation expresses the within-patient variability in relation to the grand mean of the measurements. The parameter was calculated for each model as

$$CV = \frac{SEM}{\mu} \times 100 \tag{8.11}$$

Calculation of the intra class correlation

The intra class correlation reflects the proportion of the total variability in the measurements that is attributable to true differences between the patients, and was calculated for each model as follows

$$ICC = \frac{Var_{patients}}{Var_{patients} + Var_{glands}} \tag{8.12}$$

Calculation of the median rate ratio

The median rate ratio reflects the median relative difference in the focus score between a randomly selected gland from a patient with an overall higher foci rate (i.e., a worse health status), and a randomly selected gland from a patient with an overall lower foci rate (i.e., a better health status). The parameter was estimated as

$$MRR = e^{\left(\sqrt{2\sigma_u^2}\ \Phi^{-1}(0.75)\right)}, \tag{8.13}$$

where Φ^{-1} represents the inverse of the standard normal cumulative distribution, and σ_u^2 represents the estimate of the between-individual variance of the log-rate obtained from the Poisson model.

8.6. Results

The baseline characteristics of the 32 patients recruited in the OASIS study are presented in Table 8.2. The patients were on average 54.40 (SD= 13.54) years old, with a mean BMI of 28.51 (SD=6.09). The majority of the recruited patients were females (30/32, 93.8%). The number of glands observed within the biopsies varied from 1 (1/32, 3%) to 16 (1/32, 3%), with the majority of biopsies

containing 3 to 6 glands (24/32, 75%). The distribution of both the number of foci observed within glands and the area of the glands (displayed in Figure 8.2) was positively skewed. The mean number of foci observed within the glands was 1.79 with variance 5.41 [median=1, (Q1=0, Q3=3)], while the glands were on average 3.99mm^2 large with variance 12.89mm^2 [median=3.28, (Q1=1.02, Q3=5.73)]. The corresponding mean focus score (i.e., mean number of foci per 4mm^2 of glandular area) was (1.79/3.99) × 4 =1.79. Figure 8.3 displays the focus score of all individual glands per patient, illustrating meaningful differences in the observed focus score at both the patient and gland within-patient levels. The results obtained from the three regression models are presented in Table 8.3.

Figure 8.2. Distribution of the number of foci observed within glands and the area of the glands.

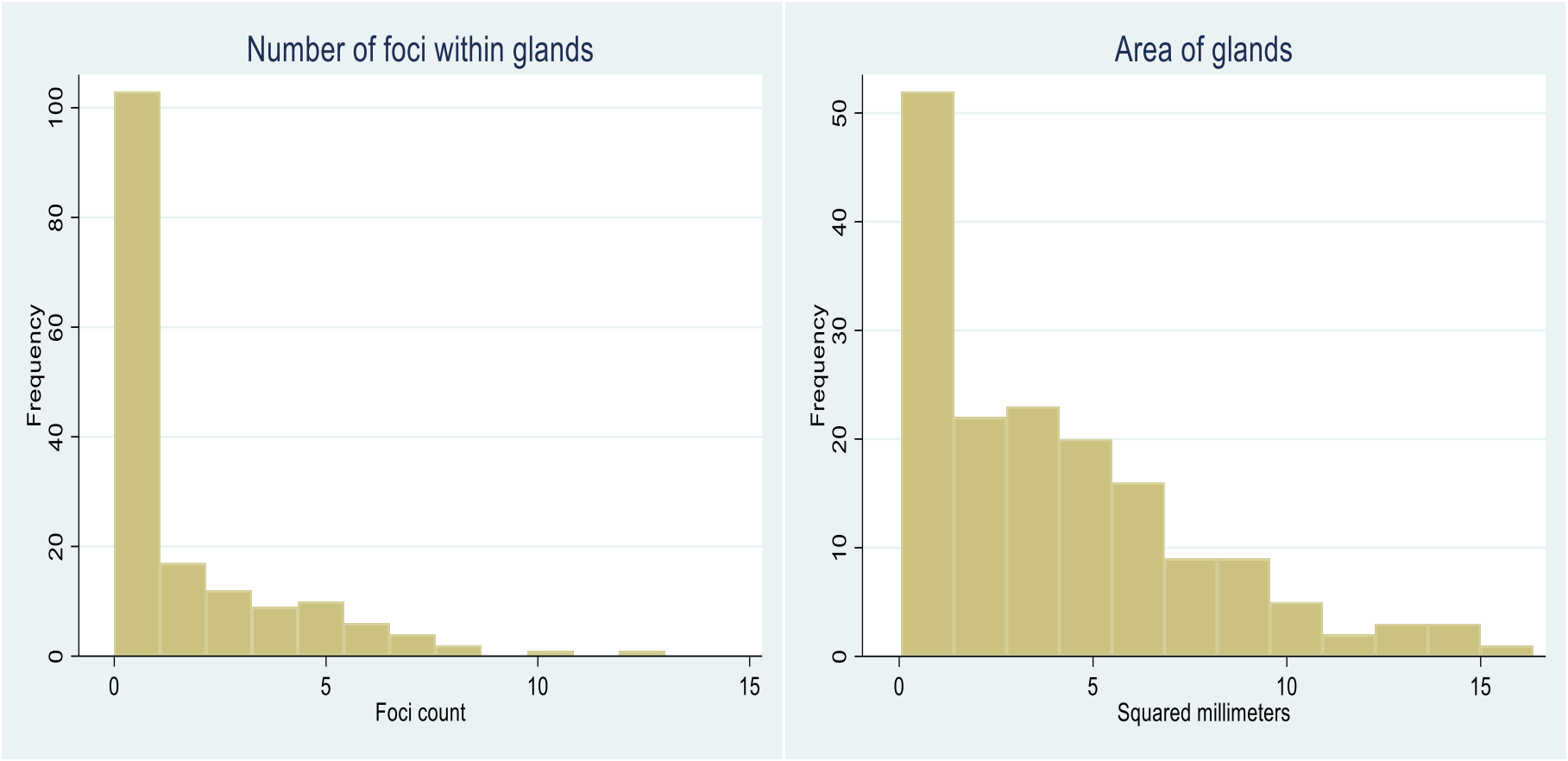
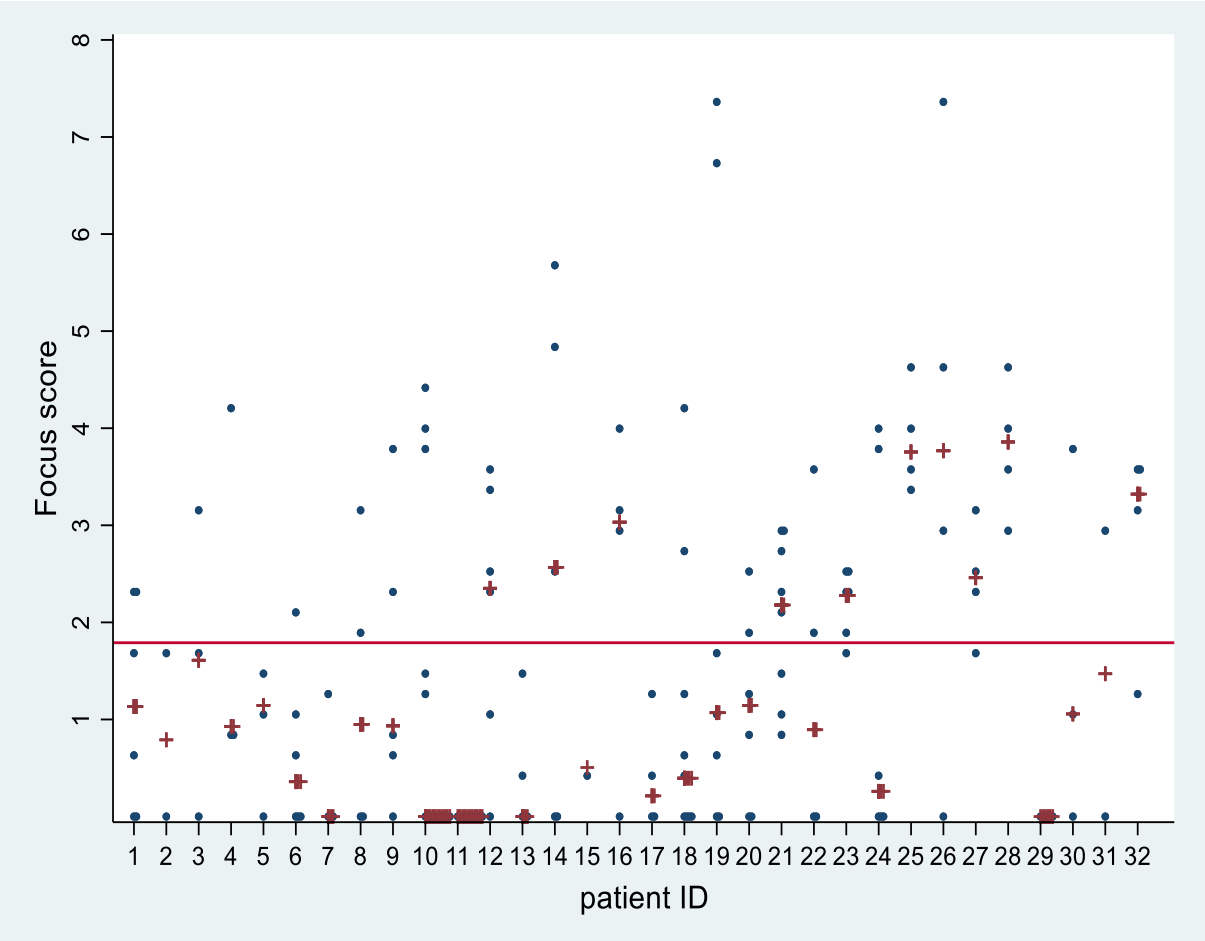


Table 8.2. Characteristics of patients recruited in the OASIS study (N=32).

- **Patient-level Characteristics**	
Age, years (mean, SD)	54.40 (13.54)
Gender [N, (%)]	30 (93.8%)
BMI (mean, SD)	28.51 (6.09)
- **Gland-level characteristics**	
Glandular area in mm^2, median [IQR]	3.28 [1.02, 5.73]
Number of glands per patient, n (%)	
- 1	1 (3)
- 2	2 (6)
- 3	4 (13)
- 4	10 (31)
- 5	3 (9)
- 6	7 (22)
- 7-16	5 (16)
Number of foci per gland, n (%)	
- 0	67 (41)
- 1	36 (21)
- 2	17 (10)
- 3	12 (7)
- 4	9 (5)
- 5	10 (6)
- 6	6 (4)
- 7	4 (2)
- 8-13	4 (2)

Figure 8.3. Dot plot of the observed focus score of each gland, split by patient. A blue dot represents the focus score of each gland. A red cross represents the median value within patient. The red line represents the mean focus score (=1.79).



8.6.1. Results from two-level Poisson model

The estimated regression intercept and between-patient variance for the log-foci rate were -0.92 [95% CI: (-1.14, -0.70)] and 0.24 [95% CI: (0.10, 0.54)], respectively. From equation 8.3, the resulting mean focus score was equal to 1.79. By the equality assumption of the Poisson distribution, the between-gland within-patient variance was also expected to be 1.79, which implies a standard error of measurement of SEM=1.34 (equation 8.10), and in turn a coefficient of variation of (1.34/1.79) × 100 = 75.0% (equation 8.11).

Using equation 8.4, the remainder variance attributed to between-patient variability in the focus score was estimated to be 0.85. This implies an ICC estimate of 0.85/ (0.85 + 1.79) = 0.323, which

denotes that 32.3% of the total variability in the focus score is attributed to differences between patients, with the remaining 67.7% attributed to differences between the glands within-patients.

Finally, the median rate ratio was estimated to be 1.59 [95% CI: (1.35, 2.01)]. This denotes that the focus score of a gland randomly selected from a patient with an overall higher foci rate (i.e., a worse health status) is expected to be at least 59% higher than the focus score of a gland randomly selected from a patient with an overall lower foci rate (i.e., a better health status), in half such comparisons performed.

8.6.2. Results from two-level negative binomial model

The estimate for the overdispersion parameter produced from the negative binomial model appeared negligible [0.02, 95% CI: (<0.01, 29.22)] and was not statistically significant based on the likelihood ratio test (p-value=0.2). As expected in such case, the results obtained from the negative binomial model were very similar to those obtained from the Poisson model (see Table 8.3).

8.6.3. Results from two-level linear model

The estimated intercept from the two-level linear regression model was 1.58 [95% CI: (1.22, 1.95)], which also corresponds to the estimate of the mean focus score. The gland-level and patient-level variance were estimated to be 2.21 [95% CI: (1.74, 2.81)] and 0.63 [95% CI: (0.27, 1.497)], respectively.

These estimates imply a standard error of measurement of 1.49, a coefficient of variation of (1.49/1.58) × 100 = 94.0%, and an intra class correlation of 0.63/ (0.63 + 2.21) = 23.2%.

8.6.4. Model comparison

The Poisson and negative binomial models produced very similar estimates, while a lower between-patient and a higher between-gland within-patient variance estimate was obtained from the linear regression model, compared to the two count-based models. The AIC and BIC values were significantly higher for the linear regression model, indicating that the analysis of the focus score as a count-measurement was more appropriate, compared to a continuous score. When analysing the focus score as a count outcome, the Poisson random effects model appeared to be a better fit for the data (AIC=463.63, BIC=469.84), compared to a negative binomial model (AIC= 465.55, BIC= 474.87).

Table 8.3. Results obtained from the analysis of the OASIS cohort (N=32).

	Two-level Poisson model	**Two-level negative binomial model**	**Two-level linear regression model**
- ***Model-based parameter estimates***			
Regression intercept, μ (95% CI)	-0.92[1] [(-1.14, -0.70)]	-0.92[1] (-1.14, -0.70)	1.58 (1.22, 1.95)
Between-patient variance, σ_u (95% CI)	0.24[1] [(0.10, 0.54)	0.23[1] (0.10, 0.55)	0.63 (0.27, 1.49)
Between-gland variance, σ_e (95% CI)	-	-	2.21 (1.74, 2.81)
Overdispersion, α (95% CI)	-	0.02 [(<0.01, 29.22), 0.2]	-
- ***Fit statistics***			
Akaike information criterion (AIC)	463.63	465.55	632.69
Bayesian information criterion (BIC)	469.84	474.87	642.01
- ***Mean and variance components***			

Grand mean	1.79 (1.50, 2.06)	1.79 (1.50, 2.06)	1.58 (1.22, 1.95)
Patient-level variance	0.85 (0.29, 1.58)	0.83 (0.25, 1.55)	0.63 (0.27, 1.49)
Gland-level variance	1.79 (1.50, 2.06)	1.86 (1.55, 2.10)	2.21 (1.74, 2.81)
- ***Parameters of reliability and measurement error***			
SEM (95% CI)	1.34 (1.22, 1.44)	1.36 (1.25, 1.48)	2.21 (1.74, 2.81)
CV (95% CI)	0.75 (0.70, 0.82)	0.76 (0.72, 0.83)	0.94 (0.78, 1.25)
ICC (%) (95% CI)	32.3 (12.1, 46.8)	31.0 (11.5, 45.6)	23.2 (4.8, 34.0)
MRR (95% CI)	1.59 (1.35, 2.01]	-	-

[1]estimated for the logarithm of the foci-rate

8.7. Discussion

This study aimed to examine potential sources of variability in the focus score using statistical methods appropriate for counts, and to evaluate the performance of these methods compared to the to the standard methods used for continuous measurements. However, the sample size of the study was not based on any formal statistical calculation, while the estimates of the within-patient variance produced from each model should be interpreted with caution due to the number of glands observed within some patients (≤3 in 22% of the patients). Furthermore, the reproducibility of the focus score was in this case not possible to be evaluated, as no repeated measurements were taken from each patient.

The Poisson and negative binomial models produced very similar estimates for the between-patient and the between-gland within-patient variance. A lower between-patient and a higher between-gland within-patient variance estimate was obtained from the linear regression model, compared to the two count-based models. The AIC and BIC values were significantly higher for the linear regression model, indicating that the analysis of the focus score as a count was in this case more appropriate, compared to a continuous score. The median rate ratio produced from the Poisson model was 1.59 [95% CI: (1.35, 2.01)], indicating significant variation in the focus score from one patient to another. However, the low ICC values obtained from all three models indicated that most

of the total variability observed in the focus score was attributed to differences between the glands within the patients, while the large values obtained for the standard error of measurement and the coefficient of variation were considered an indicative of high measurement error.

Strengths, limitations, and further work

This case study indicated strong evidence of the superiority of the count-based models over the standard linear regression model when estimating the reliability and measurement error of count-based tests. However, an obvious limitation of using the count-based models includes the further calculations required for estimating the variability of different levels. This is because, in contrast to the linear model, the estimates obtained directly from the count-based models are expressed on the logarithm scale, and thus, are no longer clinically meaningful. Furthermore, providing recommendations regarding the use of count-based models over linear models requires further work using simulation. The models should be compared across a variety of simulated scenarios (e.g., different input values for the number of individuals, measurements per individual, mean and variability at different levels) in order to identify particular scenarios where the count-based models perform better on estimating different levels of variability, compared to the standard linear regression model.

8.8. Conclusion

Although able to distinguish patients with a better health outcome to those with a worse, the focus score is highly variable within patients. Compared to a linear regression model, the two count-based models appeared to perform better when estimating the variability in the focus score attributed to true differences between patients, and the variability attributed to differences between the glands within the patients. This shows the importance of properly modelling the distribution of biomarkers

to correctly estimate the potential sources of variability. Further work involves comparing the three models across different simulated scenarios, so that recommendations of the use of count-based models over linear models can be provided.

9. Evaluation of the impact of the number of glands within labial salivary gland biopsy on the reproducibility of the focus score

9.1. Introduction

Evidence suggests that the focus score based on labial salivary gland biopsy (introduced in Chapter 8) has potential as a biomarker of primary Sjogren's syndrome, and meets several criteria required for the use in medical research. According to Fisher et al [199], the focus score is highly valid (i.e., able to accurately measure what is intended to), while a high agreement was observed between clinicians when assessing both the number of foci and the focus score of the same sample [217]. Furthermore, although invasive, the procedure is well-tolerated in experienced hands [199], and there are no significant concerns regarding its safety [218]. However, a key question arises on the reproducibility of the measurements of the focus score, as each biopsy can include a different number of glands, and there is substantial between-gland within-patient variability in the number of foci per glandular area (see Chapter 8) which the calculation of the focus score does not consider.

It is important to understand the impact of the number of glands biopsied for each individual on the estimate and precision of reproducibility. If the number of glands required per individual to obtain a reproducible estimate could be derived, this would allow minimum samples to be specified for future studies. This chapter used simulation to expand the original data set presented and analysed in Chapter 8 (where no repeated measurements were performed within patients), in order to examine the impact that different numbers of glands within labial salivary gland (LSG) biopsy may have on the reproducibility of the focus score.

9.2. Aim

The aim of this chapter was to investigate the impact of different numbers of glands in LSG biopsy on the reproducibility of the focus score, and how does this change across different simulated scenarios.

9.3. Statistical methods

The evaluation of reproducibility requires two or more consecutive measurements produced from the same individual (see Chapter 1) [219]. As repeated measurements of the focus score were not performed in the OASIS study (see Chapter 8), a simulation study (i.e., computer-based experiment that involves generating data based on known probability distributions [220]) was carried out to investigate the impact of different numbers of glands observed in a labial salivary gland biopsy on the reproducibility of the focus score. The simulation was based on the assumption that patients underwent two consecutive biopsies with two focus scores obtained from each biopsy, as this was considered the maximum number of clinically feasible biopsies. The reproducibility of the focus score was expressed as the average absolute difference between the two focus scores produced within patients, with a larger difference indicating a lower reproducibility.

9.3.1. Simulation models

The generation of the focus score requires information on both the number of foci observed within the glands and the area of the individual glands observed within each patient's biopsy. For both parameters, the choice of the simulation model (and simulation inputs, see section 9.3.2) was based on the OASIS data set (analysed in Chapter 8). The models that fitted the values observed in the OASIS data best were in turn used in this chapter to simulate each parameter required for the calculation of the focus score.

Number of foci within glands

As described in Chapter 8, the number of foci observed within glands is a count-based measurement. A random effects Negative Binomial model appeared to be a more suitable fit for the OASIS data (AIC=578.30, BIC=587.62) compared to a random effects Poisson model (AIC=610.75, BIC=616.96) when analysing the number of foci within the glands independently (i.e., not as a rate, as in Chapter 8). For each of the two biopsies taken from each patient, the number of foci within glands was simulated from a negative binomial random effects model, which is expressed as:

$$\log(foci_{ij}) = beta + u_j + \omega_{ij} \quad (9.1)$$

where $foci_{ij}$ denotes the expected number of foci within the i_{th} gland $(i = 1, \dots, n_{glands})$ within the j_{th} patient $(j = 1, \dots, n_{patients})$, $beta$ is the regression intercept, $u_j \sim N(0, \sigma_u^2)$ is the patient-level random effects parameter assumed normally distributed around a mean of 0 with variance σ_u^2, and $e^{\omega_{ij}} \sim Gamma(\frac{1}{a}, a)$ is the exponentiated gland-level overdispersion parameter assumed gamma-distributed around a mean of 1 with variance a.

For the patient-level random effects parameter (u_j), each patient was assigned the same value for both biopsies, sampled from a normal distribution with zero mean and variance σ_u^2. This was to account for the fact that the two biopsies were assumed to be performed on the same patient, with no patient-level systematic differences occurring in-between the two biopsies (e.g., a change of the patient's health status).

Area of individual glands

In contrast to the number of foci, the area of the glands is a continuous measurement, expressed in squared millimeters. The area of the glands was generated in relation to the number of foci observed within the glands, as a strong positive correlation between the two parameters was

observed in the OASIS data (Spearman's ρ=0.72, p-value<0.001), indicating that larger glands are more likely to contain a higher number of foci. For each of the two biopsies performed within each patient, the area of the glands was simulated from a linear regression mixed-effects model, which is expressed as:

$$\sqrt{Area_{ij}} = \beta_0 + \beta_1 \times Foci_{ij} + v_j + e_{ij} \quad (9.2)$$

where $\sqrt{Area_{ij}}$ is the square root of the area of the i_{th} gland $(i = 1, \dots, n_{glands})$ within the j_{th} patient $(j = 1, \dots, n_{patients})$. β_0 is the regression intercept (i.e., the mean value of the square-rooted area of glands when the number of foci equals zero), β_1 is the regression slope which accounts for the correlation between the number foci within the glands and the area of the glands, $v_j \sim N(0,\ \sigma_v^2)$ is the patient-level random effects parameter assumed normally distributed around a mean of 0 with variance σ_v^2, and $e_{ij} \sim N(0,\ \sigma_e^2)$ is the gland-level random error term assumed normally distributed around a mean of 0 with variance σ_e^2. The square root function was used as the distribution of the area of the glands within the patients recruited in the OASIS study was positively skewed, and this transformation appeared to be the most effective for normalising the data compared to any other applied, such as the log-transformation.

Similar to the number of foci within glands, the patient-level random effects parameter (v_j) took the same value for both biopsies performed on the same patient, sampled from a normal distribution with zero mean and variance σ_v^2. Again, this was to account for the fact that the two biopsies were performed on the same patient, and that there were no systematic differences in the area of the two samples taken from the patient.

9.3.2. Simulation inputs

The simulation required an input value for the following parameters: the number of patients undergoing biopsy, the number of glands within the LSG biopsy of each patient, and the regression parameters required for the generation of the number of foci within the glands, and the area of the individual glands (see equations 9.1 and 9.2). The regression parameters required for the generation of the number of foci within the glands and the area of the individual glands (see equations 9.1 and 9.2) were derived from the analysis of the OASIS data. Information on each parameter, including the notation, the estimation method (if applicable), and the input values used, is presented in Table 9.1.

Number of patients

For the base-case scenario, data sets of 30 patients were generated. This number is in agreement with the number of patients in the OASIS cohort and reflects the majority of the sample sizes seen in already conducted research studies in primary Sjogren's syndrome [221, 222]. Alterations to the base-case scenario included changing the input value of the sample size to 15 and 60 patients. These numbers reflect the range of the sample sizes seen in already conducted research studies in primary Sjogren's syndrome [221, 222].

Number of glands within LSG biopsy

The number of glands per patient assessed were 2, 3, 4, 5, 6 and 7. Each number of glands was assessed across 2500 simulated data sets (see section 9.3.4 for justification of the number of simulations), giving 15000 generated data sets in total. The range of 2 to 7 glands was based on the suggestions observed in a Delphi process conducted amongst 39 experts regarding issues around the standardisation of the glandular tissue requirements for the performance of a labial salivary gland biopsy [210].

Number of foci within glands

For the regression parameters required for the generation of number of foci within the glands (see equation 9.1), the base-case scenario input values were equal to the estimates obtained from the analysis of the OASIS data; $beta$=0.45, σ_u^2=0.64, and α=0.55. The mean number of foci within glands was estimated through the values of the regression intercept and the between-patient variance of the log number of foci, as

$$\overline{Foci_\iota} = e^{beta + \frac{\sigma_u^2}{2}} \tag{9.3}$$

Thus, the corresponding input value for the mean number of foci within glands was 2.16. From this base-case scenario, variations to the input values of the regression intercept $(beta)$, the between-patient variance of the log number of foci (σ_u^2) , and the variance of the exponentiated overdispersion parameter (a) were made one-at-a-time, as follows:

- Changing the input value of the regression intercept to 0.10 and 0.80.
- Changing the input value of the patient-level variance in the log number of foci to 0.30 and 1.32.
- Changing the input value of the variance of the exponentiated gland-level overdispersion parameter to 0.30 and 1.

The choice of these values based on the uncertainty (lower and upper 95% confidence bounds) observed in the results obtained from the analysis of the OASIS data.

Area of individual glands

For the square root of the area of the individual glands (see equation 9.2), the base-case scenario input values were equal to the estimates obtained from the analysis of the OASIS data; the

regression intercept (β_0) was 1.35, the regression slope (β_1) was 0.25, the patient-level variance (σ_v^2) was 0.09, and the gland-level variance (σ_e^2) was 0.42.

The mean area of the individual glands can then be estimated using

$$\overline{\sqrt{Area_\iota}} = \beta_0 + \beta_1 \times \overline{Foci_\iota} \tag{9.4}$$

were $\overline{Foci_\iota}$ is estimated from equation 9.3. Thus, the corresponding input value for the mean of the square root of area of the glands was 1.89mm^2. From this base-case scenario, variations to the input values of the regression intercept (β_0), the regression slope (β_1), the between-patient variance (σ_v^2), and the between-gland variance of the square root of the area of the individual glands (σ_e^2) were made one-at-a-time, as follows:

- Changing the input value of the regression intercept to 1.20 and 1.50.
- Changing the input value of the regression slope to 0.20 and 0.30.
- Changing the input value of the patient-level variance of the square root of the area of the glands to 0.02 and 0.20.
- Changing the input value of the gland-level variance or the square root of the area of the glands to 0.30 and 0.56.

These values were chosen based on the uncertainty (lower and upper 95% confidence bounds) observed in the results obtained from the analysis of the OASIS data.

Table 9.1. Notation description and estimation method for simulation parameters.

Description	Notation	Formula	Input values[2]
Sample size			
Number of patients	$n_{patients}$	-	15, **30**, 60
Number of glands within patients	n_{glands}	-	(2,3,4,5,6,7)[3]
Number of foci within glands			
Regression intercept[1]	$beta$	-	0.10, **0.45**, 0.80
Patient-level variance[1]	σ_u^2	-	0.30, **0.64**, 1.32
Gland-level variance of the exponentiated overdispersion parameter	a	-	0.30, **0.55**, 1
Mean value	$\overline{Foci_i}$	$e^{(beta + \frac{\sigma_u^2}{2})}$	-
(Square root of) area of individual glands within LSG biopsy			
Regression intercept	β_0	-	1.20, **1.35**, 1.50
Regression slope implying correlation with number of foci within glands	β_1	-	0.20, **0.25**, 0.30
Patient-level variance	σ_u^2	-	0.02, **0.09**, 0.20
Gland-level variance	σ_e^2	-	0.30, **0.42**, 0.56
Mean value	$\overline{\sqrt{Size_i}}$	$\beta_0 + \beta_1 \times \overline{Foci_i}$	-

[1]estimated on the logarithm scale, [2]value in bold indicates the input value used for the base-case scenario,[3]all 6 input values were used in all different analyses performed to examine how reproducibility changes for different simulated numbers of glands.

9.3.3. Analysis of generated data

2500 data sets were simulated (see section 9.3.4 for justification of the number of simulations) for each of the six numbers of glands assessed, with each data set containing a value for the number of foci and the square root of the area of each gland, for both biopsies performed on each patient. The mean values of the two parameters ($\overline{Foci_i}$ and $\overline{\sqrt{Size}_i}$) were then calculated for each data set. These values were assessed for bias, which was calculated as the absolute difference between the average of the 2500 generated estimates, expressed as mean (SD) or median (IQR), and the expected input values for each parameter (estimated from equations 9.3 and 9.4).

The two focus scores for each biopsy performed on each patient were then calculated as the total number of foci observed within each biopsy, over the total glandular area, multiplied by $4mm^2$ [210].The reproducibility of the focus score was expressed for each patient as the absolute difference between the two focus scores, which was summarised within each simulated data set using the median and interquartile range (IQR). The 2500 estimates produced for the median absolute difference and IQR were in turn summarised using mean (SD) or median (IQR), as appropriate.

The data generation and analysis were then repeated for a different number of glands observed in the biopsy, and the results were compared across the six simulated numbers of glands (2 to 7). Visual assessments of how the estimated average absolute difference and its corresponding precision (i.e., the IQR) changes for a unit increase in the number of glands were made using boxplots, displaying the median of the produced 2500 estimates along with the lower and upper quartiles for each simulated number of glands, and scatterplots, plotting the produced median absolute difference within each data set against the associated interquartile range.

9.3.4. Justification of the number of simulations

The median absolute difference between the two focus scores produced within the patients, which reflects the reproducibility of the focus score, was considered the main parameter of interest. The number of simulations for each different number of glands was determined using the following formula:

$$Monte\ Carlo\ SE = \frac{SD}{\sqrt{n_{sim}}}, \tag{9.5}$$

where $Monte\ Carlo\ SE$ and SD are Monte Carlo standard error and standard deviation of the median absolute difference between the two focus scores, across n_{sim} generated data sets. A Monte Carlo standard error of ≤0.5% will allow the median absolute difference between the two focus scores to be estimated with a satisfactory degree of precision. The chosen value of 0.5% was equal to the value used in the example presented in Morris et al [220].

To obtain an estimate of the standard deviation of the median absolute difference between the two measurements, an initial small simulation of 200 replications was performed for each of the 6 glands, assuming 30 patients within each simulated data set. The base-case scenario was used for the remaining data generation inputs. The standard deviation of the median absolute difference varied from 0.11 to 0.23 across the 6 simulated numbers of glands, with higher values observed for lower numbers of glands per patient. With 2 glands within biopsy, the upper 95% confidence bound was equal to 0.26. For the remaining 5 scenarios, the upper 95% confidence bound did not exceed 0.25. Thus, a conservative estimate of 0.25 was used to determine the number of replications required, in order to obtain a Monte Carlo standard error of 0.5%. From equation 9.5, this implies that 2500 replications were required.

9.4. Results

9.4.1 Base-case scenario

Generation of gland-level parameters

The 2500 estimates produced for the mean number of foci within glands and the mean of the square root of the area of glands were normally distributed for both simulated biopsies and all 6 simulated numbers of glands within biopsy (Figures 9.1 and 9.2). If present, the bias of the mean number of foci within glands was no more than an absolute value of 0.02, while no bias was noted for the mean of the square root of the area of the glands, across all simulated numbers of glands within each biopsy (Table 9.2).

Reproducibility of the focus score

The 2500 estimates produced for the median absolute difference between the two within-patient focus scores and the corresponding interquartile range were normally distributed for all 6 simulated numbers of glands (Figures 9.4 and 9.5). The standard deviation of the median absolute difference was lower than or equal to the pre-specified value of 0.25, allowing the parameter to be estimated with satisfactory precision across all simulated numbers of glands.

A decrease in the median absolute difference was noted for every unit increase in the number of glands; the mean of the 2500 estimates produced for the median absolute difference was 1.05 (SD=0.25) for 2 glands, which was reduced to 0.52 (SD=0.12) for 7 glands. A similar trend was observed for the interquartile range, with the mean of 2500 IQR estimates reducing from 1.60 (SD=0.48) for 2 glands, down to 0.71 (SD=0.18) for 7 glands (see Table 9.3).

Table 9.2. Generation of gland-level parameters – base case scenario.

Input values – Sample size		**Input values – Number of foci within glands**				**Input values – Square root of the area of glands**					**Results – 1st LSG biopsy**		**Results – 2nd LSG biopsy**	
$n_{patients}$	n_{glands}	$beta$	σ_u^2	$alpha$	$\overline{Foci_\iota}$	β_0	β_1	σ_v^2	σ_e^2	$\overline{\sqrt{Area_\iota}}$	$\widehat{\overline{Foci_\iota}}$	$\widehat{\overline{\sqrt{Area_\iota}}}$	$\widehat{\overline{Foci_\iota}}$	$\widehat{\overline{\sqrt{Area_\iota}}}$
30	2	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.18 (0.50)	1.89 (0.16)	2.17 (0.51)	1.89 (0.16)
30	3	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.45)	1.89 (0.14)	2.15 (0.44)	1.89 (0.14)
30	4	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.45)	1.89 (0.14)	2.16 (0.45)	1.89 (0.14)
30	5	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.42)	1.89 (0.13)	2.16 (0.43)	1.89 (0.13)
30	6	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.42)	1.89 (0.13)	2.16 (0.42)	1.89 (0.13)
30	7	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.42)	1.89 (0.12)	2.17 (0.42)	1.89 (0.13)

Table 9.3. Reproducibility of the focus score – base case scenario.

$n_{patients}$	n_{glands}	**Median of absolute difference in focus score (mean, SD)**	**IQR of absolute difference in focus score (mean, SD)**
30	2	1.05 (0.25)	1.60 (0.48)
30	3	0.82 (0.19)	1.19 (0.33)
30	4	0.71 (0.16)	0.98 (0.28)
30	5	0.62 (0.14)	0.86 (0.22)
30	6	0.56 (0.13)	0.77 (0.20)
30	7	0.52 (0.12)	0.71 (0.18)

Figure 9.1. Distribution of the produced 2500 estimates for the mean of the number of foci within glands and each LSG biopsy – base case scenario.

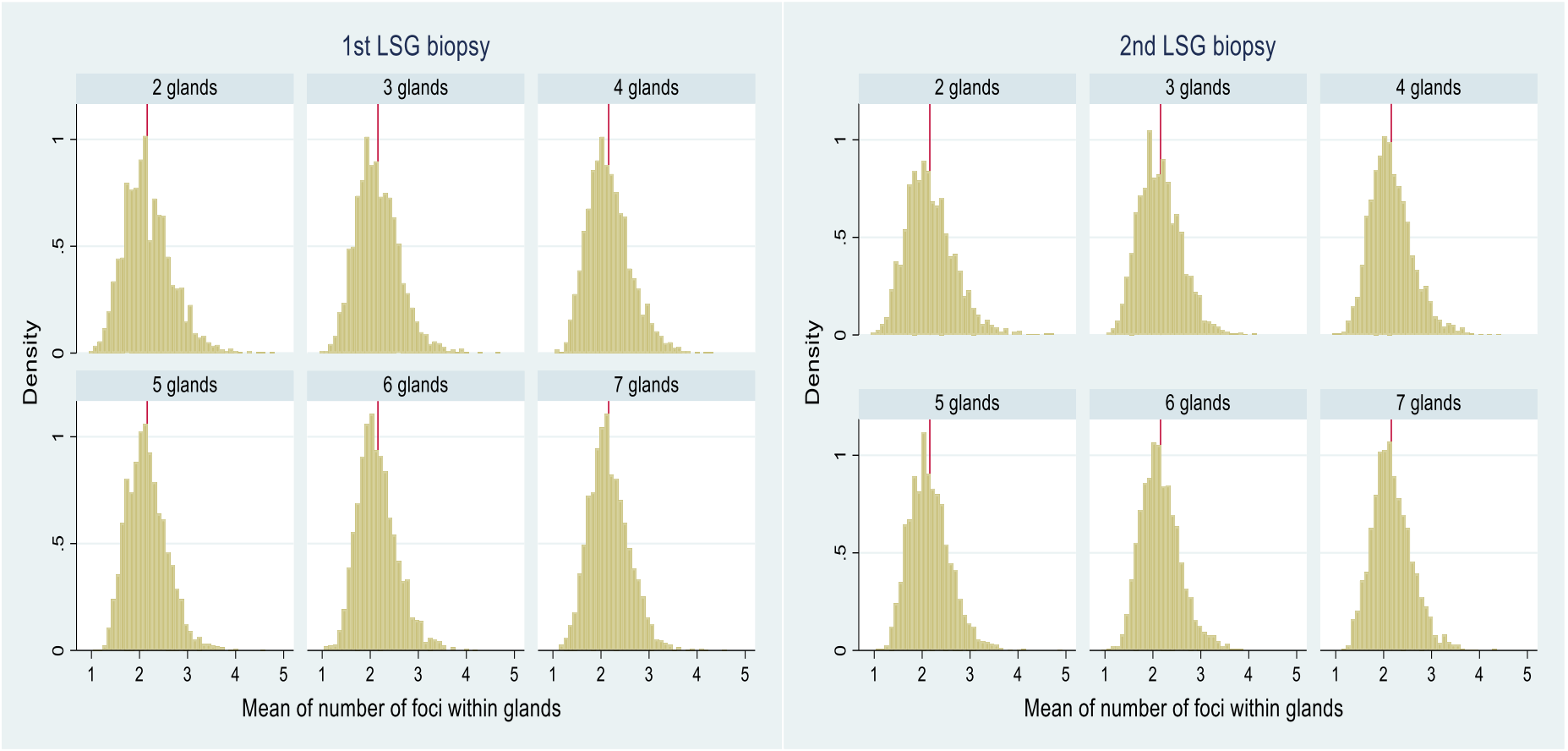


Figure 9.2. Distribution of the produced 2500 estimates for the mean of the square root of the area of the glands and each LSG biopsy – base case scenario.

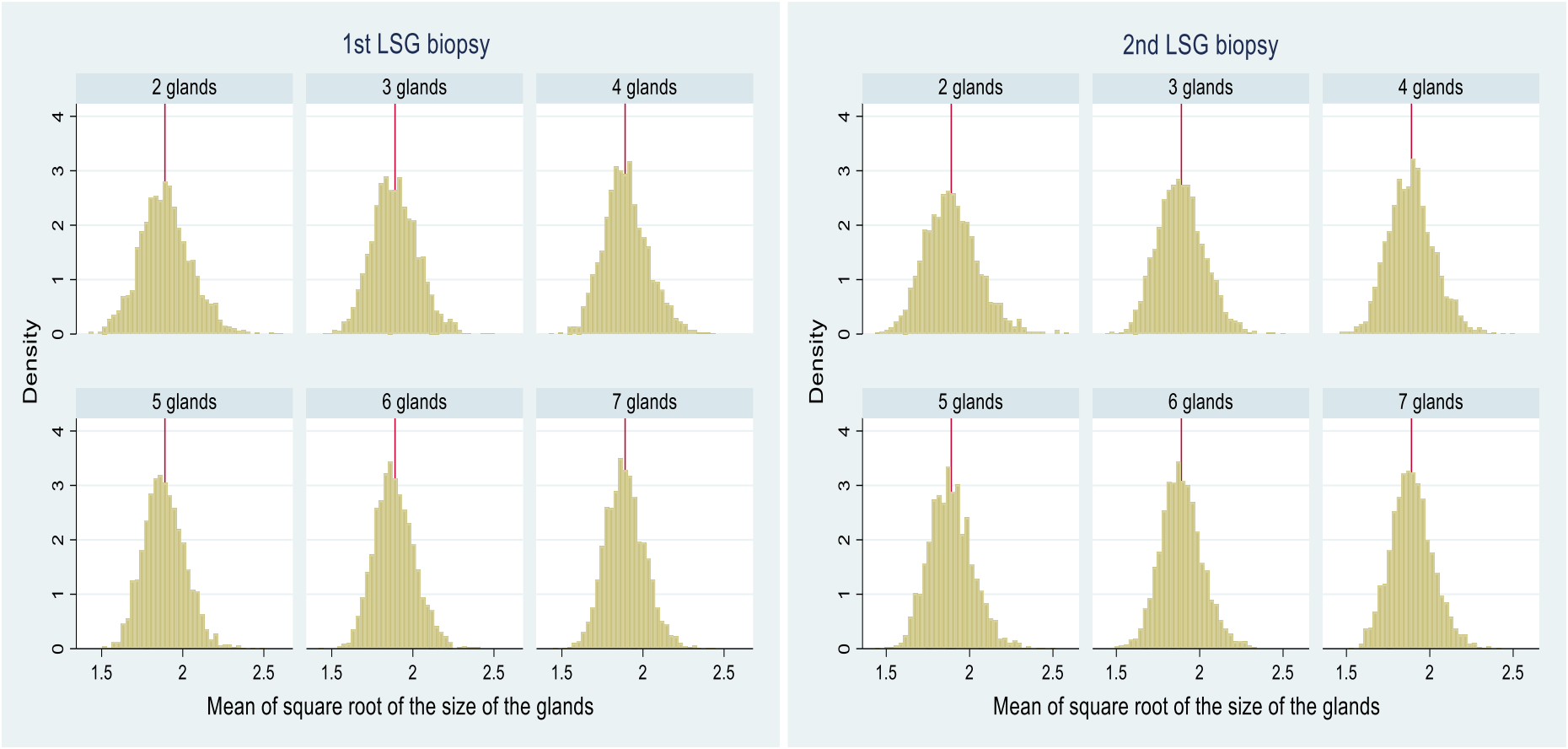


Figure 9.3. Scatterplot of the 2500 median differences in the focus score against the associated interquartile range – base case scenario. The reference value of median=0.5 is used to aid comparissons between different simulated numbers of glands.

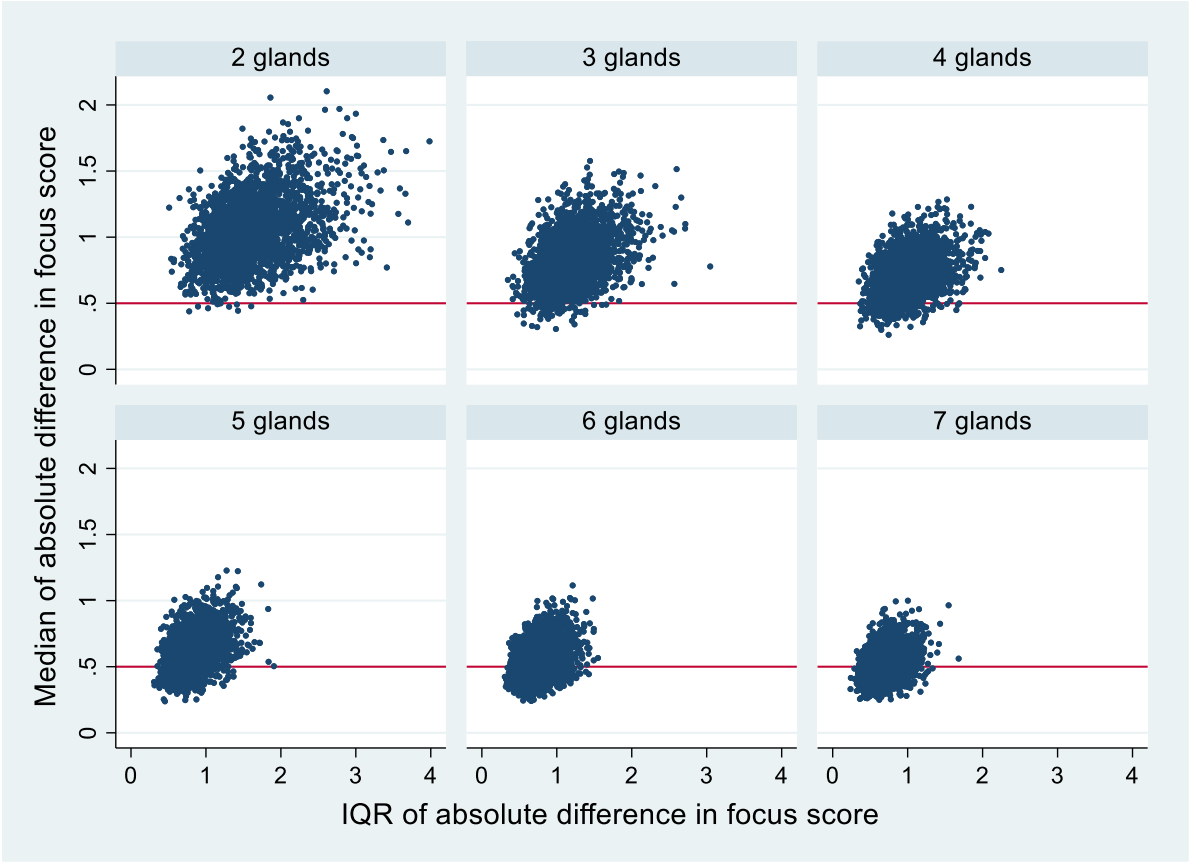


Figure 9.4. Boxplot of median absolute difference in focus score for each simulated number of glands – base case scenario. The reference value of median=0.5 is used to aid comparissons between different simulated numbers of glands.

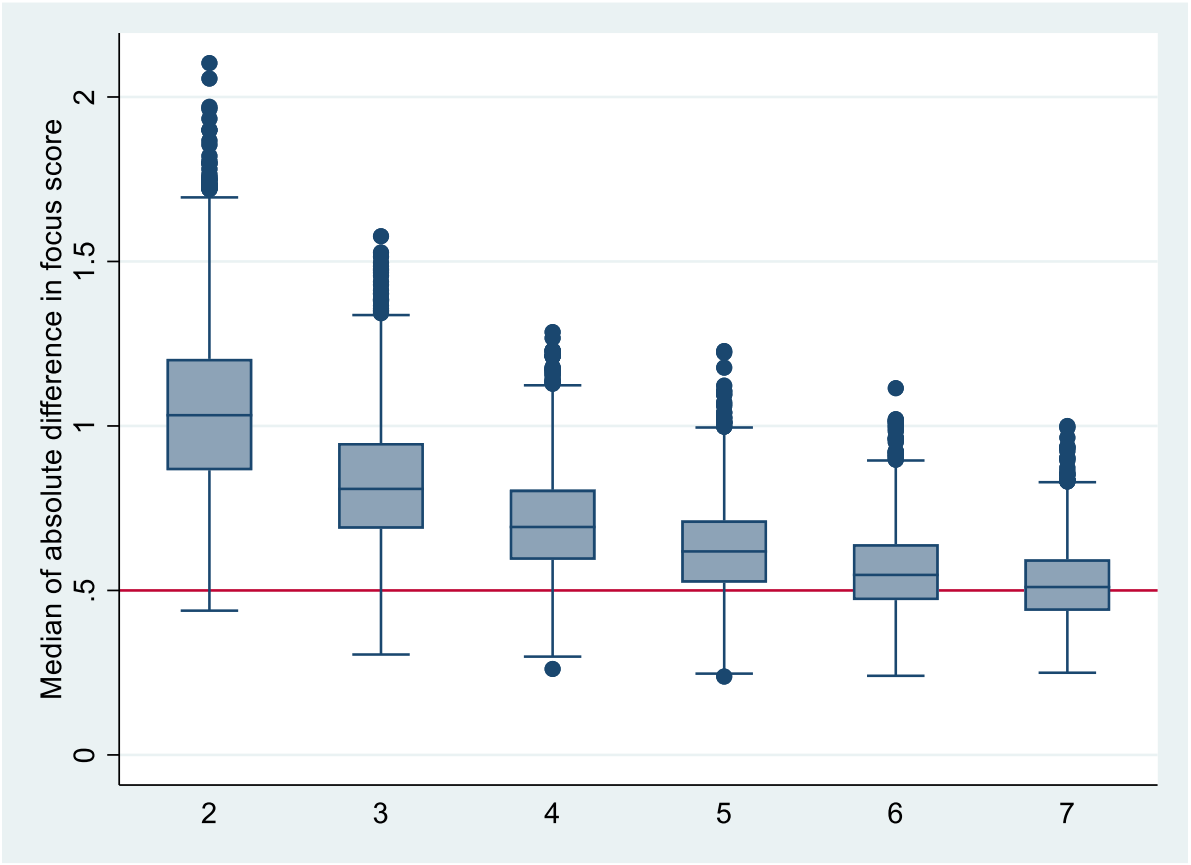
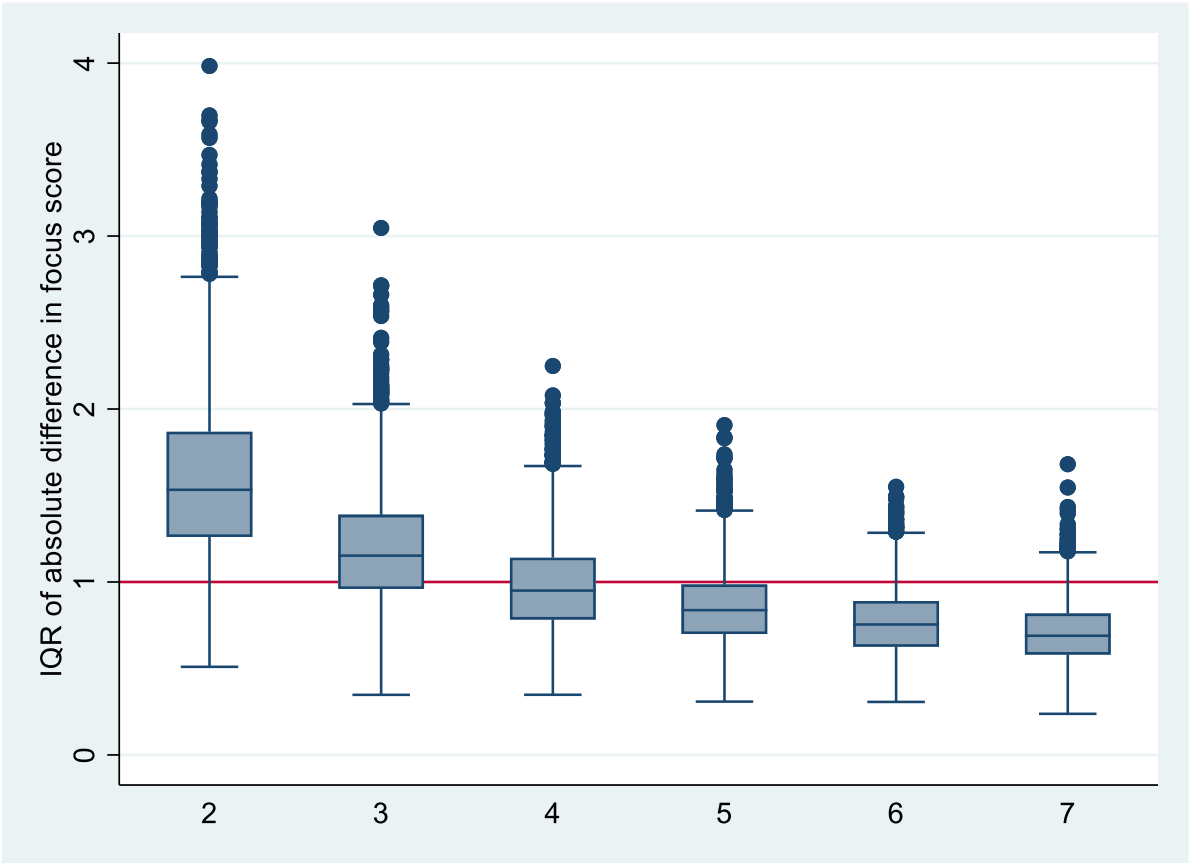


Figure 9.5. Boxplot of IQR of absolute difference in focus score for each simulated number of glands – base case scenario. The reference value of IQR=1 is used to aid comparissons between different simulated numbers of glands.



9.4.2. Sensitivity analysis

Generation of gland-level parameters

The bias for the mean number of foci and square root of the area of the glands remained similar to the base-case scenario, when altering the input value of the sample size, or any of the regression parameters required for the generation of the square root of the area of the glands (see equations 9.1 and 9.2). For both the mean number of foci and square root of the area of the glands, the produced estimates were biased by no more than an absolute value of 0.02 and 0.01, respectively, across all simulated scenarios (Tables 9.4, 9.6 and 9.8).

Reproducibility of the focus score

The results obtained for all different input values remained similar to the base-case scenario, with the mean of 2500 estimates produced for the median absolute difference and interquartile range decreasing for a unit increase in the number of glands (Tables 9.5, 9.7 and 9.9).

A higher input value for σ_u^2 (i.e., a higher between-patient variability in the number of foci), β_0 (i.e., a larger glandular area), and β_1 (i.e., a stronger positive correlation between the number of foci within glands and the area of the glands) produced lower estimates for the median absolute difference.

Lower IQR estimates (i.e., a higher precision within the generated data sets) were observed for a higher number of patients, a higher input value for β_0 and β_1, and a lower input value for σ_v^2 (i.e., a lower between-patient variability in the glandular area) and σ_e^2 (i.e., a lower between-gland within-patient variability in the glandular area).

Table 9.4. Generation of gland-level parameters – varying the sample size.

Input values – Sample size		**Input values – Number of foci within glands**				**Input values – Square root of the area of glands**					**Results – 1st LSG biopsy**		**Results – 2nd LSG biopsy**	
$n_{patients}$	n_{glands}	$beta$	σ_u^2	a	$\overline{Foci_\iota}$	β_0	β_1	σ_v^2	σ_e^2	$\overline{\sqrt{Area_\iota}}$	$\widehat{\overline{Foci_\iota}}$	$\widehat{\overline{\sqrt{Area_\iota}}}$	$\widehat{\overline{Foci_\iota}}$	$\widehat{\overline{\sqrt{Area_\iota}}}$
15	2	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.72)	1.89 (0.23)	2.14 (0.69)	1.89 (0.23)
15	3	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.66)	1.89 (0.21)	2.15 (0.67)	1.89 (0.21)
15	4	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.14 (0.61)	1.89 (0.19)	2.17 (0.62)	1.89 (0.19)
15	5	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.60)	1.89 (0.18)	2.17 (0.61)	1.89 (0.19)
15	6	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.18 (0.59)	1.90 (0.18)	2.18 (0.60)	1.90 (0.18)
15	7	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.58)	1.89 (0.18)	2.16 (0.59)	1.89 (0.18)
60	2	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.36)	1.89 (0.11)	2.16 (0.36)	1.89 (0.11)
60	3	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.34)	1.89 (0.10)	2.16 (0.33)	1.89 (0.10)
60	4	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.32)	1.89 (0.10)	2.16 (0.32)	1.89 (0.10)
60	5	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.30)	1.89 (0.09)	2.16 (0.30)	1.89 (0.09)
60	6	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.31)	1.89 (0.09)	2.16 (0.31)	1.89 (0.09)
60	7	0.45	0.64	0.55	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.29)	1.89 (0.09)	2.17 (0.29)	1.89 (0.09)

Table 9.5. Reproducibility of the focus score – varying the sample size.

$n_{patients}$	n_{glands}	**Median of absolute difference in focus score (mean, SD)**	**IQR of absolute difference in focus score (mean, SD)**
15	2	1.08 (0.37)	1.79 (0.80)
15	3	0.84 (0.29)	1.31 (0.52)
15	4	0.70 (0.23)	1.05 (0.40)
15	5	0.63 (0.20)	0.92 (0.35)
15	6	0.57 (0.18)	0.83 (0.29)
15	7	0.52 (0.17)	0.76 (0.28)
60	2	1.05 (0.18)	1.59 (0.32)
60	3	0.82 (0.13)	1.19 (0.24)
60	4	0.70 (0.12)	0.99 (0.19)
60	5	0.62 (0.10)	0.86 (0.16)
60	6	0.56 (0.09)	0.77 (0.14)
60	7	0.51 (0.08)	0.71 (0.13)

Table 9.6. Generation of gland-level parameters – varying the regression parameters required for the number of foci within glands.

Input values – Sample size		**Input values – Number of foci within glands**				**Input values – Square root of the area of glands**					**Results – 1st LSG biopsy**		**Results – 2nd LSG biopsy**	
$n_{patients}$	n_{glands}	$beta$	σ_u^2	a	$\overline{Foci_i}$	β_0	β_1	σ_v^2	σ_e^2	$\overline{\sqrt{Area_i}}$	$\widehat{\overline{Foci_i}}^1$	$\widehat{\overline{\sqrt{Area_i}}}^1$	$\widehat{\overline{Foci_i}}^1$	$\widehat{\overline{\sqrt{Area_i}}}^1$
- ***Varying the regression intercept*** $(\boldsymbol{beta})$														
30	2	**0.10**	0.64	0.55	1.52	1.35	0.25	0.09	0.42	1.73	1.52 (0.37)	1.74 (0.14)	1.52 (0.37)	1.73 (0.14)
30	3	**0.10**	0.64	0.55	1.52	1.35	0.25	0.09	0.42	1.73	1.52 (0.34)	1.73 (0.12)	1.53 (0.34)	1.73 (0.12)
30	4	**0.10**	0.64	0.55	1.52	1.35	0.25	0.09	0.42	1.73	1.51 (0.31)	1.73 (0.11)	1.52 (0.31)	1.73 (0.11)
30	5	**0.10**	0.64	0.55	1.52	1.35	0.25	0.09	0.42	1.73	1.53 (0.31)	1.73 (0.11)	1.52 (0.31)	1.73 (0.11)
30	6	**0.10**	0.64	0.55	1.52	1.35	0.25	0.09	0.42	1.73	1.52 (0.29)	1.73 (0.10)	1.52 (0.30)	1.73 (0.10)
30	7	**0.10**	0.64	0.55	1.52	1.35	0.25	0.09	0.42	1.73	1.53 (0.31)	1.73 (0.10)	1.53 (0.30)	1.73 (0.10)
30	2	**0.80**	0.64	0.55	3.06	1.35	0.25	0.09	0.42	2.12	3.06 (0.68)	2.12 (0.19)	3.06 (0.71)	2.12 (0.20)
30	3	**0.80**	0.64	0.55	3.06	1.35	0.25	0.09	0.42	2.12	3.07 (0.68)	2.12 (0.19)	3.05 (0.67)	2.11 (0.19)
30	4	**0.80**	0.64	0.55	3.06	1.35	0.25	0.09	0.42	2.12	3.07 (0.63)	2.12 (0.18)	3.06 (0.65)	2.12 (0.18)
30	5	**0.80**	0.64	0.55	3.06	1.35	0.25	0.09	0.42	2.12	3.07 (0.60)	2.12 (0.17)	3.07 (0.61)	2.12 (0.17)
30	6	**0.80**	0.64	0.55	3.06	1.35	0.25	0.09	0.42	2.12	3.06 (0.58)	2.12 (0.16)	3.05 (0.59)	2.11 (0.17)
30	7	**0.80**	0.64	0.55	3.06	1.35	0.25	0.09	0.42	2.12	3.06 (0.57)	2.12 (0.16)	3.06 (0.58)	2.12 (0.16)
- ***Varying the patient-level variance of the log-number of foci*** $(\boldsymbol{\sigma_u^2})$														
30	2	0.45	**0.30**	0.55	1.82	1.35	0.25	0.09	0.42	1.81	1.83 (0.33)	1.81 (0.13)	1.83 (0.35)	1.81 (0.13)
30	3	0.45	**0.30**	0.55	1.82	1.35	0.25	0.09	0.42	1.81	1.82 (0.29)	1.80 (0.12)	1.83 (0.29)	1.81 (0.11)
30	4	0.45	**0.30**	0.55	1.82	1.35	0.25	0.09	0.42	1.81	1.82 (0.27)	1.81 (0.11)	1.82 (0.27)	1.81 (0.11)
30	5	0.45	**0.30**	0.55	1.82	1.35	0.25	0.09	0.42	1.81	1.83 (0.26)	1.81 (0.10)	1.82 (0.27)	1.81 (0.10)
30	6	0.45	**0.30**	0.55	1.82	1.35	0.25	0.09	0.42	1.81	1.81 (0.25)	1.80 (0.10)	1.82 (0.25)	1.80 (0.10)
30	7	0.45	**0.30**	0.55	1.82	1.35	0.25	0.09	0.42	1.81	1.83 (0.25)	1.81 (0.09)	1.83 (0.24)	1.81 (0.09)
30	2	0.45	**1.32**	0.55	3.04	1.35	0.25	0.09	0.42	2.11	3.04 (1.11)	2.11 (0.29)	3.02 (1.05)	2.10 (0.28)
30	3	0.45	**1.32**	0.55	3.04	1.35	0.25	0.09	0.42	2.11	3.04 (1.04)	2.11 (0.27)	3.03 (1.07)	2.11 (0.28)
30	4	0.45	**1.32**	0.55	3.04	1.35	0.25	0.09	0.42	2.11	3.04 (1.03)	2.11 (0.27)	3.05 (1.05)	2.11 (0.27)
30	5	0.45	**1.32**	0.55	3.04	1.35	0.25	0.09	0.42	2.11	3.04 (1.00)	2.11 (0.26)	3.02 (0.99)	2.11 (0.26)
30	6	0.45	**1.32**	0.55	3.04	1.35	0.25	0.09	0.42	2.11	3.05 (0.99)	2.11 (0.26)	3.04 (1.02)	2.11 (0.27)
30	7	0.45	**1.32**	0.55	3.04	1.35	0.25	0.09	0.42	2.11	3.02 (0.92)	2.11 (0.24)	3.03 (0.94)	2.11 (0.25)
- ***Varying the gland-level variance of the exponentiated overdispersion parameter*** $(\boldsymbol{a})$														

30	2	0.45	0.64	**0.30**	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.47)	1.89 (0.15)	2.16 (0.47)	1.89 (0.15)
30	3	0.45	0.64	**0.30**	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.44)	1.89 (0.14)	2.15 (0.44)	1.89 (0.14)
30	4	0.45	0.64	**0.30**	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.42)	1.89 (0.13)	2.17 (0.43)	1.89 (0.14)
30	5	0.45	0.64	**0.30**	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.41)	1.89 (0.13)	2.17 (0.41)	1.89 (0.13)
30	6	0.45	0.64	**0.30**	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.40)	1.89 (0.12)	2.15 (0.41)	1.89 (0.13)
30	7	0.45	0.64	**0.30**	2.16	1.35	0.25	0.09	0.42	1.89	2.16 (0.41)	1.89 (0.13)	2.15 (0.41)	1.89 (0.12)
30	2	0.45	0.64	**1**	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.57)	1.89 (0.18)	2.16 (0.56)	1.89 (0.17)
30	3	0.45	0.64	**1**	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.52)	1.89 (0.16)	2.16 (0.50)	1.89 (0.15)
30	4	0.45	0.64	**1**	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.49)	1.89 (0.14)	2.16 (0.48)	1.89 (0.14)
30	5	0.45	0.64	**1**	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.45)	1.89 (0.14)	2.15 (0.45)	1.89 (0.14)
30	6	0.45	0.64	**1**	2.16	1.35	0.25	0.09	0.42	1.89	2.15 (0.44)	1.89 (0.13)	2.15 (0.45)	1.89 (0.13)
30	7	0.45	0.64	**1**	2.16	1.35	0.25	0.09	0.42	1.89	2.17 (0.45)	1.89 (0.13)	2.16 (0.44)	1.89 (0.13)

Table 9.7. Reproducibility of the focus score – varying the regression parameters required for the number of foci within glands.

$n_{patients}$	n_{glands}	$beta$	σ_u^2	a	Median of absolute difference in focus score (mean, SD)	IQR of absolute difference in focus score (mean, SD)
-	***Varying the regression intercept*** $(beta)$					
30	2	**0.10**	0.64	0.55	1.01 (0.25)	1.59 (0.47)
30	3	**0.10**	0.64	0.55	0.81 (0.19)	1.18 (0.33)
30	4	**0.10**	0.64	0.55	0.68 (0.16)	0.97 (0.27)
30	5	**0.10**	0.64	0.55	0.60 (0.14)	0.85 (0.23)
30	6	**0.10**	0.64	0.55	0.54 (0.13)	0.75 (0.20)
30	7	**0.10**	0.64	0.55	0.50 (0.11)	0.70 (0.18)
30	2	**0.80**	0.64	0.55	1.06 (0.25)	1.58 (0.47)
30	3	**0.80**	0.64	0.55	0.82 (0.19)	1.16 (0.33)
30	4	**0.80**	0.64	0.55	0.70 (0.16)	0.97 (0.27)
30	5	**0.80**	0.64	0.55	0.62 (0.14)	0.84 (0.22)
30	6	**0.80**	0.64	0.55	0.55 (0.12)	0.75 (0.19)
30	7	**0.80**	0.64	0.55	0.51 (0.12)	0.69 (0.18)
-	***Varying the patient-level variance of the log-number of foci*** (σ_u^2)					
30	2	0.45	**0.30**	0.55	1.13 (0.27)	1.67 (0.51)
30	3	0.45	**0.30**	0.55	0.88 (0.20)	1.25 (0.34)
30	4	0.45	**0.30**	0.55	0.74 (0.17)	1.02 (0.28)
30	5	0.45	**0.30**	0.55	0.66 (0.14)	0.90 (0.24)
30	6	0.45	**0.30**	0.55	0.60 (0.13)	0.81 (0.21)
30	7	0.45	**0.30**	0.55	0.55 (0.12)	0.74 (0.19)
30	2	0.45	**1.32**	0.55	0.94 (0.24)	1.52 (0.45)
30	3	0.45	**1.32**	0.55	0.75 (0.17)	1.12 (0.32)
30	4	0.45	**1.32**	0.55	0.64 (0.15)	0.92 (0.26)
30	5	0.45	**1.32**	0.55	0.57 (0.13)	0.80 (0.21)
30	6	0.45	**1.32**	0.55	0.51 (0.12)	0.72 (0.19)
30	7	0.45	**1.32**	0.55	0.47 (0.11)	0.65 (0.17)
-	***Varying the gland-level variance of the exponentiated overdispersion parameter*** (a)					
30	2	0.45	0.64	**0.30**	1.03 (0.25)	1.59 (0.49)

30	3	0.45	0.64	**0.30**	0.81 (0.19)	1.16 (0.32)
30	4	0.45	0.64	**0.30**	0.68 (0.16)	0.97 (0.27)
30	5	0.45	0.64	**0.30**	0.61 (0.14)	0.84 (0.23)
30	6	0.45	0.64	**0.30**	0.55 (0.13)	0.76 (0.20)
30	7	0.45	0.64	**0.30**	0.51 (0.11)	0.69 (0.18)
30	2	0.45	0.64	**1**	1.09 (0.26)	1.64 (0.48)
30	3	0.45	0.64	**1**	0.86 (0.20)	1.22 (0.33)
30	4	0.45	0.64	**1**	0.73 (0.16)	1.01 (0.26)
30	5	0.45	0.64	**1**	0.65 (0.14)	0.89 (0.23)
30	6	0.45	0.64	**1**	0.58 (0.13)	0.78 (0.20)
30	7	0.45	0.64	**1**	0.53 (0.12)	0.73 (0.19)

Table 9.8. Generation of gland-level parameters – varying the regression parameters required for the square root of the area of the glands.

Input values – Sample size		**Input values – Number of foci within glands**				**Input values – Square root of the area of glands**					**Results – 1st LSG biopsy**		**Results – 2nd LSG biopsy**	
$n_{patients}$	n_{glands}	$beta$	σ_u^2	a	$\overline{Foci_i}$	β_0	β_1	σ_v^2	σ_e^2	$\overline{\sqrt{Area_i}}$	$\widehat{\overline{Foci_i}}^1$	$\widehat{\overline{\sqrt{Area_i}}}^1$	$\widehat{\overline{Foci_i}}^1$	$\widehat{\overline{\sqrt{Area_i}}}^1$
- ***Varying the regression intercept*** $(\boldsymbol{\beta_0})$														
30	2	0.45	0.64	0.55	2.16	**1.20**	0.25	0.09	0.42	1.74	2.15 (0.52)	1.74 (0.16)	2.16 (0.49)	1.73 (0.16)
30	3	0.45	0.64	0.55	2.16	**1.20**	0.25	0.09	0.42	1.74	2.16 (0.47)	1.74 (0.15)	2.15 (0.46)	1.74 (0.15)
30	4	0.45	0.64	0.55	2.16	**1.20**	0.25	0.09	0.42	1.74	2.16 (0.44)	1.74 (0.14)	2.16 (0.44)	1.74 (0.14)
30	5	0.45	0.64	0.55	2.16	**1.20**	0.25	0.09	0.42	1.74	2.16 (0.43)	1.74 (0.13)	2.16 (0.43)	1.74 (0.13)
30	6	0.45	0.64	0.55	2.16	**1.20**	0.25	0.09	0.42	1.74	2.16 (0.41)	1.74 (0.12)	2.16 (0.41)	1.74 (0.13)
30	7	0.45	0.64	0.55	2.16	**1.20**	0.25	0.09	0.42	1.74	2.15 (0.42)	1.74 (0.12)	2.15 (0.42)	1.74 (0.12)
30	2	0.45	0.64	0.55	2.16	**1.50**	0.25	0.09	0.42	2.04	2.16 (0.52)	2.04 (0.16)	2.14 (0.50)	2.04 (0.16)
30	3	0.45	0.64	0.55	2.16	**1.50**	0.25	0.09	0.42	2.04	2.15 (0.47)	2.04 (0.15)	2.14 (0.47)	2.04 (0.15)
30	4	0.45	0.64	0.55	2.16	**1.50**	0.25	0.09	0.42	2.04	2.17 (0.45)	2.04 (0.14)	2.17 (0.46)	2.04 (0.14)
30	5	0.45	0.64	0.55	2.16	**1.50**	0.25	0.09	0.42	2.04	2.17 (0.44)	2.04 (0.14)	2.17 (0.42)	2.04 (0.13)
30	6	0.45	0.64	0.55	2.16	**1.50**	0.25	0.09	0.42	2.04	2.18 (0.42)	2.04 (0.13)	2.17 (0.43)	2.04 (0.13)
30	7	0.45	0.64	0.55	2.16	**1.50**	0.25	0.09	0.42	2.04	2.16 (0.41)	2.04 (0.12)	2.15 (0.40)	2.04 (0.12)
- ***Varying the regression slope*** $(\boldsymbol{\beta_1})$														
30	2	0.45	0.64	0.55	2.16	1.35	**0.20**	0.09	0.42	1.78	2.15 (0.52)	1.78 (0.15)	2.16 (0.51)	1.78 (0.14)
30	3	0.45	0.64	0.55	2.16	1.35	**0.20**	0.09	0.42	1.78	2.18 (0.48)	1.78 (0.13)	2.18 (0.48)	1.78 (0.13)
30	4	0.45	0.64	0.55	2.16	1.35	**0.20**	0.09	0.42	1.78	2.16 (0.44)	1.78 (0.12)	2.15 (0.44)	1.78 (0.12)
30	5	0.45	0.64	0.55	2.16	1.35	**0.20**	0.09	0.42	1.78	2.15 (0.43)	1.78 (0.11)	2.16 (0.43)	1.78 (0.11)
30	6	0.45	0.64	0.55	2.16	1.35	**0.20**	0.09	0.42	1.78	2.17 (0.41)	1.79 (0.11)	2.17 (0.41)	1.78 (0.11)
30	7	0.45	0.64	0.55	2.16	1.35	**0.20**	0.09	0.42	1.78	2.14 (0.42)	1.78 (0.11)	2.15 (0.41)	1.78 (0.11)
30	2	0.45	0.64	0.55	2.16	1.35	**0.30**	0.09	0.42	2	2.15 (0.50)	1.99 (0.18)	2.16 (0.49)	2.00 (0.18)
30	3	0.45	0.64	0.55	2.16	1.35	**0.30**	0.09	0.42	2	2.15 (0.46)	2.00 (0.16)	2.16 (0.47)	2.00 (0.17)
30	4	0.45	0.64	0.55	2.16	1.35	**0.30**	0.09	0.42	2	2.16 (0.44)	1.99 (0.15)	2.15 (0.44)	1.99 (0.15)
30	5	0.45	0.64	0.55	2.16	1.35	**0.30**	0.09	0.42	2	2.17 (0.43)	2.00 (0.15)	2.17 (0.44)	2.00 (0.15)
30	6	0.45	0.64	0.55	2.16	1.35	**0.30**	0.09	0.42	2	2.16 (0.42)	2.00 (0.14)	2.15 (0.42)	2.00 (0.15)
30	7	0.45	0.64	0.55	2.16	1.35	**0.30**	0.09	0.42	2	2.15 (0.42)	2.00 (0.15)	2.15 (0.42)	2.00 (0.15)
- ***Varying the patient-level variance of the square root of the area of the glands*** $(\boldsymbol{\sigma_v^2})$														

30	2	0.45	0.64	0.55	2.16	1.35	0.25	**0.02**	0.42	1.89	2.14 (0.51)	1.88 (0.16)	2.15 (0.52)	1.89 (0.15)
30	3	0.45	0.64	0.55	2.16	1.35	0.25	**0.02**	0.42	1.89	2.18 (0.47)	1.90 (0.14)	2.17 (0.46)	1.89 (0.14)
30	4	0.45	0.64	0.55	2.16	1.35	0.25	**0.02**	0.42	1.89	2.16 (0.44)	1.89 (0.13)	2.17 (0.46)	1.89 (0.13)
30	5	0.45	0.64	0.55	2.16	1.35	0.25	**0.02**	0.42	1.89	2.16 (0.43)	1.89 (0.12)	2.16 (0.43)	1.89 (0.12)
30	6	0.45	0.64	0.55	2.16	1.35	0.25	**0.02**	0.42	1.89	2.16 (0.41)	1.89 (0.12)	2.16 (0.43)	1.89 (0.12)
30	7	0.45	0.64	0.55	2.16	1.35	0.25	**0.02**	0.42	1.89	2.17 (0.43)	1.89 (0.12)	2.17 (0.42)	1.89 (0.12)
30	2	0.45	0.64	0.55	2.16	1.35	0.25	**0.20**	0.42	1.89	2.15 (0.51)	1.89 (0.17)	2.16 (0.51)	1.89 (0.17)
30	3	0.45	0.64	0.55	2.16	1.35	0.25	**0.20**	0.42	1.89	2.17 (0.46)	1.89 (0.16)	2.18 (0.47)	1.89 (0.16)
30	4	0.45	0.64	0.55	2.16	1.35	0.25	**0.20**	0.42	1.89	2.17 (0.44)	1.89 (0.15)	2.17 (0.45)	1.89 (0.15)
30	5	0.45	0.64	0.55	2.16	1.35	0.25	**0.20**	0.42	1.89	2.16 (0.41)	1.89 (0.14)	2.15 (0.42)	1.89 (0.15)
30	6	0.45	0.64	0.55	2.16	1.35	0.25	**0.20**	0.42	1.89	2.16 (0.42)	1.89 (0.14)	2.16 (0.43)	1.89 (0.14)
30	7	0.45	0.64	0.55	2.16	1.35	0.25	**0.20**	0.42	1.89	2.16 (0.41)	1.89 (0.14)	2.16 (0.41)	1.89 (0.14)
-	***Varying the gland-level variance of the square root of the area of the glands*** $(\boldsymbol{\sigma_e^2})$													
30	2	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.30**	1.89	2.14 (0.50)	1.88 (0.15)	2.17 (0.52)	1.89 (0.16)
30	3	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.30**	1.89	2.15 (0.46)	1.89 (0.14)	2.16 (0.47)	1.89 (0.14)
30	4	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.30**	1.89	2.16 (0.46)	1.89 (0.14)	2.15 (0.45)	1.89 (0.14)
30	5	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.30**	1.89	2.15 (0.41)	1.89 (0.12)	2.16 (0.42)	1.89 (0.13)
30	6	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.30**	1.89	2.16 (0.41)	1.89 (0.13)	2.15 (0.42)	1.89 (0.12)
30	7	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.30**	1.89	2.16 (0.42)	1.89 (0.13)	2.16 (0.42)	1.89 (0.12)
30	2	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.56**	1.89	2.16 (0.51)	1.89 (0.17)	2.15 (0.49)	1.89 (0.17)
30	3	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.56**	1.89	2.16 (0.46)	1.89 (0.15)	2.15 (0.46)	1.89 (0.15)
30	4	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.56**	1.89	2.16 (0.44)	1.89 (0.14)	2.16 (0.44)	1.89 (0.14)
30	5	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.56**	1.89	2.16 (0.43)	1.89 (0.14)	2.17 (0.43)	1.89 (0.13)
30	6	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.56**	1.89	2.15 (0.42)	1.89 (0.13)	2.16 (0.42)	1.89 (0.13)
30	7	0.45	0.64	0.55	2.16	1.35	0.25	0.09	**0.56**	1.89	2.17 (0.41)	1.89 (0.13)	2.17 (0.42)	1.89 (0.13)

Table 9.9. Reproducibility of the focus score – varying the regression parameters required for the square root of the area of the glands.

$n_{patients}$	n_{glands}	β_0	β_1	σ_v^2	σ_e^2	Median of absolute difference in focus score (mean, SD)	IQR of absolute difference in focus score (mean, SD)
- ***Varying the regression intercept*** (β_0)							
30	2	**1.20**	0.25	0.09	0.42	1.27 (0.31)	2.02 (0.65)
30	3	**1.20**	0.25	0.09	0.42	0.99 (0.24)	1.47 (0.44)
30	4	**1.20**	0.25	0.09	0.42	0.84 (0.19)	1.20 (0.34)
30	5	**1.20**	0.25	0.09	0.42	0.74 (0.17)	1.05 (0.29)
30	6	**1.20**	0.25	0.09	0.42	0.66 (0.15)	0.93 (0.25)
30	7	**1.20**	0.25	0.09	0.42	0.61 (0.14)	0.84 (0.22)
30	2	**1.50**	0.25	0.09	0.42	0.89 (0.21)	1.30 (0.37)
30	3	**1.50**	0.25	0.09	0.42	0.70 (0.16)	0.99 (0.27)
30	4	**1.50**	0.25	0.09	0.42	0.60 (0.14)	0.82 (0.22)
30	5	**1.50**	0.25	0.09	0.42	0.53 (0.12)	0.72 (0.19)
30	6	**1.50**	0.25	0.09	0.42	0.48 (0.11)	0.65 (0.17)
30	7	**1.50**	0.25	0.09	0.42	0.45 (0.10)	0.59 (0.15)
- ***Varying the regression slope*** (β_1)							
30	2	1.35	**0.20**	0.09	0.42	1.26 (0.31)	1.97 (0.61)
30	3	1.35	**0.20**	0.09	0.42	0.99 (0.23)	1.46 (0.42)
30	4	1.35	**0.20**	0.09	0.42	0.84 (0.20)	1.20 (0.34)
30	5	1.35	**0.20**	0.09	0.42	0.74 (0.17)	1.04 (0.28)
30	6	1.35	**0.20**	0.09	0.42	0.67 (0.16)	0.94 (0.25)
30	7	1.35	**0.20**	0.09	0.42	0.62 (0.14)	0.86 (0.23)
30	2	1.35	**0.30**	0.09	0.42	0.92 (0.22)	1.35 (0.40)
30	3	1.35	**0.30**	0.09	0.42	0.71 (0.16)	1.00 (0.29)
30	4	1.35	**0.30**	0.09	0.42	0.60 (0.13)	0.83 (0.22)
30	5	1.35	**0.30**	0.09	0.42	0.53 (0.12)	0.73 (0.19)
30	6	1.35	**0.30**	0.09	0.42	0.48 (0.11)	0.65 (0.17)
30	7	1.35	**0.30**	0.09	0.42	0.44 (0.10)	0.59 (0.15)
- ***Varying the patient-level variance of the square root of the area of the glands*** (σ_v^2)							
30	2	1.35	0.25	**0.02**	0.42	1.06 (0.24)	1.50 (0.43)

30	3	1.35	0.25	**0.02**	0.42	0.82 (0.18)	1.12 (0.29)
30	4	1.35	0.25	**0.02**	0.42	0.70 (0.16)	0.93 (0.23)
30	5	1.35	0.25	**0.02**	0.42	0.62 (0.13)	0.81 (0.20)
30	6	1.35	0.25	**0.02**	0.42	0.56 (0.12)	0.73 (0.18)
30	7	1.35	0.25	**0.02**	0.42	0.52 (0.11)	0.67 (0.16)
30	2	1.35	0.25	**0.20**	0.42	1.06 (0.27)	1.77 (0.61)
30	3	1.35	0.25	**0.20**	0.42	0.83 (0.21)	1.32 (0.41)
30	4	1.35	0.25	**0.20**	0.42	0.71 (0.17)	1.08 (0.33)
30	5	1.35	0.25	**0.20**	0.42	0.63 (0.15)	0.94 (0.28)
30	6	1.35	0.25	**0.20**	0.42	0.57 (0.14)	0.84 (0.25)
30	7	1.35	0.25	**0.20**	0.42	0.52 (0.12)	0.78 (0.23)
-	***Varying the gland-level variance of the square root of the area of the glands*** $(\boldsymbol{\sigma_e^2})$						
30	2	1.35	0.25	0.09	**0.30**	0.99 (0.23)	1.43 (0.40)
30	3	1.35	0.25	0.09	**0.30**	0.78 (0.18)	1.09 (0.30)
30	4	1.35	0.25	0.09	**0.30**	0.67 (0.15)	0.91 (0.24)
30	5	1.35	0.25	0.09	**0.30**	0.59 (0.13)	0.80 (0.21)
30	6	1.35	0.25	0.09	**0.30**	0.53 (0.12)	0.72 (0.19)
30	7	1.35	0.25	0.09	**0.30**	0.49 (0.11)	0.66 (0.17)
30	2	1.35	0.25	0.09	**0.56**	1.11 (0.28)	1.79 (0.59)
30	3	1.35	0.25	0.09	**0.56**	0.87 (0.21)	1.29 (0.38)
30	4	1.35	0.25	0.09	**0.56**	0.74 (0.17)	1.05 (0.29)
30	5	1.35	0.25	0.09	**0.56**	0.65 (0.15)	0.91 (0.25)
30	6	1.35	0.25	0.09	**0.56**	0.58 (0.13)	0.81 (0.22)
30	7	1.35	0.25	0.09	**0.56**	0.54 (0.12)	0.75 (0.20)

9.5. Discussion

This chapter investigated the impact of different numbers of glands in LSG biopsy on the reproducibility of the focus score. Knowledge of this impact will help clinicians decide on the minimum number of biopsy glands required, when considering using the focus score in future studies of Sjogren's syndrome.

Simulation was used to generate a value for the number of foci and the area of each gland within LSG biopsy, accounting for potential between and within patient variability for each parameter. For the area of the glands, data generation additionally accounted for potential correlation with the number of foci observed within glands, based on the assumption that larger glands are more likely to contain a higher number of foci. This approach yielded the desired estimates for the number of foci within glands and the area of the glands, with the observed bias being negligible (or in some cases absent) across the simulated scenarios.

An alternative approach which was initially considered was to generate the foci rate (i.e., number of foci per squared millimeter of glandular area) and area of each gland, multiply the foci rate with the area of each gland to obtain the corresponding number of foci, and in turn calculate each patient's focus score. However, this approach has the disadvantage of not accounting for the potential correlation between number of foci observed within glands and the area of the glands, and was not preferred in the end.

The simulation suggested that the absolute difference between two consecutive measurements produced within patients was decreasing for a unit increase in the number of glands. This difference was very similar for a number of 6 and 7 glands, and was reduced by approximately 50% compared to a number of 2 glands. A similar downward trend was noted for the interquartile range, indicating that an increase in the number of glands allowed the absolute difference to be estimated with a higher precision within the generated data sets. Similar results were obtained for all different input values used for sensitivity analysis, with a higher reproducibility observed for a higher between-

patient variability in the number of foci, a larger glandular area, and a stronger correlation between the number of foci and the area of the glands.

9.6. Conclusion

A larger number of biopsied glands provides a better and more precise estimate of the reproducibility of the focus score. Samples containing 6 or 7 glands reduced the absolute difference between two consecutive measurements down to approximately 50% of the difference observed for samples of 2 glands. The findings aim to help clinicians decide on what the minimum number of biopsied glands should be, so that they make best use of the focus score in future research studies of Sjogren's syndrome.

10. Discussion and conclusions

Before biomarkers are used in medical research and practice, researchers and medical professionals need to be aware of the potential error that the produced measurements may be accompanied by. High measurement error may lead to false conclusions regarding the presence, deterioration, or severity of a medical condition, the effects and safety of medical treatments, and the occurrence of future clinical outcomes.

This thesis explored statistical issues around the estimation of the reliability and measurement error of biomarkers. The overarching objectives of the thesis were:

i) To propose statistical methods for the meta-analysis of parameter estimates of reliability and measurement error of continuous biomarkers, reported across primary studies.

ii) To propose statistical methods for the analysis of primary studies examining the reliability and measurement error of biomarkers expressed as counts, when the normality of measurements produced at different levels (assumed in standard methods used for of continuous measurements) is likely to be violated.

10.1. Thesis overview and summary of findings

Chapter 1 explained the key concepts used in the thesis; provided an overview of the potential impact of measurement error on medical research and practice (with examples); defined the scope of this thesis and specified the main aims.

Current approaches for the design and statistical analysis of primary studies examining the reliability and measurement error of continuous biomarkers

The current approaches used for the design and statistical analysis of primary studies examining the reliability and measurement error of continuous biomarkers were presented in detail in Chapter 2. Primary studies in the field are typically designed to recruit a group of individuals and obtain multiple measurements from each individual, at each potential level of variability. There has been a long-standing debate on which individuals should be recruited in such studies (current guidelines for laboratory based-tests recommend that only healthy individuals should be considered), while limited guidance is available regarding the number of individuals and number of within-individual measurements required for designing such studies. Prior to analysis, researchers are often concerned whether data meet the assumptions of normality and heteroscedasticity (if not, the log-transformation is applied), and whether data are free from any significant outliers.

The variability at each level is then estimated through a nested analysis of variance (ANOVA) or a linear regression random effects model. The estimates obtained from these analyses allow a number of parameters to be calculated, including: the standard error of measurement, the smallest detectable change, the intra-class correlation, the coefficient of variation, the index of individuality, and the reference change values. Alternative parameters for examining the reliability and error of continuous measurements include standard correlation coefficients (e.g., Pearson or Spearman) and the limits of agreement, respectively, while one additional parameter is available for examining potential inter/intra-observer variability in binary or ordinal responses (the Kappa coefficient).

Primary study to examine the reliability and error of grip strength measurements

The case study analysis (Chapter 3) allowed illustration of the methods for continuous measurements (introduced in Chapter 2) to examine the reliability and measurement error of the

measurements of grip strength, obtained from a digital dynamometer. The results revealed that, although not error free, digital dynamometers produce highly reliable measurements and may have potential as part of evaluating loss of strength in patients with sarcopenia and chronic inflammatory disease.

Current systematic review and meta-analytic methods used for examining test reliability and measurement error

Chapter 4 examined the current state of the review process and meta-analytic methods used in systematic reviews examining test reliability and measurement error. A methodological review of 219 systematic reviews was carried out. The quality of the review methods used in the identified systematic reviews was variable. Encouragingly, the majority of the reviews provided a clear description of the inclusion criteria and study characteristics, and searched at least one database in addition to PUBMED/MEDLINE. However, a high number of reviews did not provide a clear description of the search strategy. Screening titles and abstracts, extracting the relevant data, or assessing the quality of primary studies was often performed by a single author, while some reviews (42/219, 20%) did not even assess the quality of the included primary studies. Furthermore, the approach used for screening, data extraction, and quality assessment was often not clearly reported.

Of the identified reviews, only 22 (10%) carried out a quantitative synthesis of the data reported in primary studies, with 16 of these 22 producing a weighted average estimate of reliability and/or measurement error. The meta-analytic methods that were used for this purpose were found to have the following limitations:

- the violation of the normality assumption that standard models for meta-analysis hold, as the estimates of parameters of reliability and measurement error reported across studies are not expected to be normally distributed.

- the estimation of the within sampling variance not being independent of the reported estimate for some parameters, such as the intra class correlation, the standard error of measurement, and the standard deviation of the difference between two within-individual measurements (required for constructing the limits of agreement). This makes the inverse-variance weights (i.e., the standard approach for weighting studies in meta-analyses) no more applicable, as the weight assigned to each study will partly depend on the magnitude of the reported estimate (and not entirely on the sample size).

Whilst the use of the Fisher's Z transformation has been shown to solve these two issues when pooling estimates of correlation coefficients reported across studies, no approaches were noted for key parameters of measurement error, such as the limits of agreement, the standard error of measurement, and the coefficient of variation.

New methods proposed for the meta-analysis of key parameters of measurement error

Alternative methods for the meta-analysis of the limits of agreement, the standard error of measurement, and the coefficient of variation were introduced in Chapter 5. For the limits of agreement, the framework proposed by Tipton and Shuster [166] was presented, while two new methods were developed for the standard error of measurement and the coefficient of variation. The methods focused on satisfying the assumption of underlying normality of the reported study-level parameter estimates, as well as stabilizing the sampling error variance of the parameters.

Systematic review and meta-analysis examining the reliability and error of grip strength measurements produced from hand-held dynamometers

The methods for meta-analysis presented in Chapter 5 were in turn applied in Chapter 6, where a systematic review and meta-analysis was conducted to provide summary evidence of the reliability

and error of grip strength measurements, produced from different types of dynamometers and across different populations. All methods were very effective in normalising the sampling distribution of the reported parameters of reliability and measurement error. The results obtained were very similar to those from the case study analysis (Chapter 3), indicating the grip strength measurements from hand-held dynamometers are produced with excellent reliability and low measurement error.

Alternative methods for estimating the reliability and measurement error of count-based biomarkers

Alternative methods, potentially more appropriate for the statistical analysis of primary studies examining the reliability and measurement error of count-based biomarkers, were identified in the literature and presented in Chapter 7. The different sources of variability are this time estimated from a Poisson or a negative binomial random effects model, and the methods described in Leckie et al [202] and Austin et al [204]. Standard parameters of reliability and measurement error may in turn be calculated (i.e., the standard error of measurement, the intraclass correlation, and the coefficient of variation), while a new, potentially useful parameter expressing the reliability of count-based measurements was also introduced (i.e., the median rate ratio).

Case-study to evaluate the performance of count-based methods compared to the standard normality-based approach

The count-based methods were applied in Chapter 8 to estimate the between and within-individual variability of the focus score, using data from 32 patients with Sjogren's syndrome who underwent labial salivary gland biopsy as a case study. The performance of these methods was also compared to the standard methods used for continuous measurements. Whilst all methods indicated that the

focus score is likely to be subject to high within-individual variability (i.e., measurement error) and may not always be fit for use in clinical practice, the count-based methods provided less biased estimates of the between and within-individual variability compared to the standard method used for continuous measurements.

Simulation study to evaluate how different numbers of glands impact reproducibility of the focus score

A subsequent simulation study (Chapter 9) investigated the impact of different numbers of biopsy glands on the reproducibility of the focus score. The results illustrated that the reproducibility of the focus score was highly dependent on the number of glands within the biopsy. A unit increase in the number of glands improved the reproducibility of the focus score, with samples containing 6 or 7 glands reducing the absolute difference between two consecutive measurements down to approximately 50% of the difference observed for samples of 2 glands.

10.2. Strengths and limitations

10.2.1. Strengths

Chapter 2 provided a novel and robust evaluation of the issues around the design and statistical analysis of primary studies examining the reliability and measurement error of biomarkers. The chapter highlighted the flaws of the current guidelines for laboratory tests, particularly with respect to the target population, the assessment of normality at the within-individual level, and the detection and removal of outlies. Current gaps in the field of reliability and measurement error were also identified, including the limited guidance available for estimating the sample size required for a primary study, and the lack of statistical methods for the meta-analysis of parameters of reliability

and measurement error of continuous biomarkers, and statistical methods for primary analysis of count-measurements.

The case study reported in Chapter 3 provided strong evidence that digital dynamometers can reliably measure the grip strength of patients with sarcopenia and chronic inflammatory disease. The study was reasonably sized and analysed using the best available methods. The analysis also allowed the application of standard methods used for examining test reliability and measurement error to be clearly illustrated.

The methodological review in Chapter 4 was the first to explore the current practice for conducting and reporting systematic reviews of the reliability and measurement error of biomarkers, and the current state of statistical methods available for the meta-analysis of parameters of reliability and measurement error. The findings indicated important flaws in how such reviews are conducted and reported, and how parameter estimates of reliability and measurement error reported across primary studies are combined. In order to provide a clear overview of current practice, the review used two different databases (MEDLINE and EMBASE) for study identification, which demonstrates best practice and methodological rigour [162, 163].

Based on the limitations of the existing methods, novel statistical methods for the meta-analysis of key parameters of measurement error such as the standard error of measurement and the coefficient of variation were developed and presented in Chapter 5. The new methods were developed using well-established mathematical relationships. The performance of these methods was evaluated in a case study analysis of 80 primary studies evaluating the reliability and error of grip strength measurements (Chapter 6). The results demonstrated these methods to be very effective in normalising the distribution of the reported study-level estimates.

The systematic review and meta-analysis performed in Chapter 6 also reinforced the primary study analysis in Chapter 3, which showed that hand-held dynamometers used to evaluate grip strength produce highly reliable measurements. The systematic review was conducted using robust review

methods. Primary studies were identified by searching two electronic databases (MEDLINE and EMBASE). The study selection and quality assessment was undertaken independently by two reviewers, while the data extractions were checked by a second reviewer, with any queries being discussed and resolved by consensus. In addition, compared to a previous review examining the reliability and measurement error of hand-held dynamometers, this review provided quantitative summary evidence using the meta-analytic methods introduced in the thesis.

Chapter 7 was the first to present alternative methods for estimating the reliability and error of count measurements, where the assumption of the normality of measurements produced at different variability levels is known to often be violated. A case-study (Chapter 8) was also used to illustrate how these methods work, and to indicate the potential benefits of using these methods over the standard normality-based approach when estimating the reliability and measurement error of count-based tests.

Finally, the work presented in Chapter 9 was the first to formally investigate the impact of different numbers of glands on the reproducibility of the focus score, as any existing suggestions regarding the number of biopsy glands required for a reproducible estimate were based on opinions of experts taking part in a Delphi technique, rather than a formal statistical analysis. To make best effort to reflect real life, all simulated assumptions were based on real data, as well as in-depth discussions with clinical collaborators.

10.2.2. Limitations

In Chapters 2 and 7, the approaches used for the design and statistical analysis of primary studies examining test reliability and measurement error, and statistical methods used for estimating sources of variability of count outcomes, were identified through a literature review which was not systematic. For both chapters, the aim was to understand the available approaches and their limitations, and a systematic review was not considered an efficient way to locate this information.

In the primary study examining the reliability and error of grip strength measurements (Chapter 3), the number of visits and measurements performed within each visit were small (2 and 3 respectively), which did not allow the corresponding components of variability to be estimated with high precision. However, these numbers are reflective of how the test is often be used in practice (e.g., elderly/frail participants would not be expected to attend more than two visits or produce more than three measurements in a session). Although the number of patients recruited in the study (N=84) was relatively large compared to similarly designed studies identified for the systematic review in Chapter 6 (median sample size=35 [Q1=25, Q3=76]), this number was not based on a formal sample size calculation. Furthermore, while the results obtained are only applicable to patients with sarcopenia and chronic conditions, and cannot be generalised to a wider population, for an elderly and often frail population it may be particularly important to obtain population-specific estimates.

In Chapter 4, the selection of systematic reviews and data extraction was undertaken by a single reviewer, and it is advised that these two tasks should be performed independently by at least two reviewers. However, this was a methodologic review aiming to provide a comprehensive reflection of current practice as opposed to producing unbiased estimates of e.g., accuracy or effectiveness, and hence did not need to be as comprehensive in study identification. Furthermore, the review did not consider the adequacy of tools used for assessing the quality of the primary studies included in the identified reviews, while the overall quality of the reviews was not assessed using a formal checklist (e.g., similar to AMSTAR 2 [165], which is intended for systematic reviews of healthcare interventions). The latter were outside the scope of this work, which aimed to provide a general overview of current practice in this under-researched area, and primarily to identify statistical approaches for the synthesis of the data reported in primary studies.

With respect to the meta-analysis of parameters of reliability and measurement error, all methods proposed were based on transforming the estimates reported in primary studies, in order to account

for the non-normal sampling distribution of the parameters. The use of these transformations makes the calculation of a weighted average estimate more complex. For example, obtaining a summary estimate for the standard error of measurement involves squaring and log-transforming the study-level estimates prior to meta-analysis, then taking the square root of the exponential of the produced weighted average estimate. However, all transformations were shown to satisfy the assumption of normality of the study-level estimates, which is important when using standard meta-analysis models (e.g., the DerSimonian and Laird random effects model).

In Chapter 6, high heterogeneity in the estimates reported across primary studies was noted in most of the meta-analyses performed. This heterogeneity was to some extent expected given the broad criteria used for study inclusion. Not identifying any factors causing this heterogeneity, and particularly not assessing the effect of the different patient populations included, can be considered a limitation of this study.

Chapter 7 presented alternative statistical models for primary analysis, potentially more appropriate for estimating sources of variability of count-based tests. In contrast to the standard normality-based model, these models do not produce the variability estimates on the original measurement scale, but the logarithmic, which does not allow direct interpretation of the estimates. This can be considered a limitation, given that additional (and computationally intensive) calculations are required to revert the estimates to the original scale. However, the case study analysis (Chapter 8) showed that these methods provide less biased estimates of reliability and measurement error compared to the standard normality-based approach, when evaluating a count-based test.

In Chapter 8, the number of 32 patients recruited in the study was not based on a formal sample size calculation, while the small numbers of patients and glands observed within some patients (≤3 in 22% of the patients) did not allow the models to estimate the two components of variability with high precision. Furthermore, the reproducibility of the focus score using this data set was not

possible to be evaluated, as no repeated measurements were taken from each patient. The thesis aimed to tackle this issue by expanding this data set using simulation (Chapter 9).

In Chapter 9, the generation of the focus score required information on the number of foci observed within the glands and the size of the individual glands observed within each patient's biopsy. For both parameters, the choice of the simulating distribution and input values was based on a real data set of patients with primary Sjogren's syndrome, who undergone labial salivary gland biopsy (i.e., the OASIS cohort, used as a case study in Chapter 8). However, the data set used was relatively small, and it is possible that different distributional patterns would have been observed if a different (larger) data set was used as a guide, leading to different decisions regarding the simulation of the two parameters. This can be considered a general issue when simulating data. To tackle this issue as much as possible, the simulation considered multiple different scenarios based on the 95% certainty around the parameters estimates observed in the analysis of the original data, and assessed how the obtained results differed across these scenarios.

10.3. Implications for medical research and practice

This thesis highlighted that current guidelines available for the design of primary studies examining the reliability and measurement error of medical tests have flaws and need updating. The target population should not be restricted to healthy individuals, as the variability of a test may change between healthy and diseased populations. Therefore, testing healthy individuals may not provide any information on the reliability and measurement error of tests intended for a diseased population. Statistical tests for assessing whether within-individual measurements deviate from normality should be used with caution (if used at all). This is because primary studies of test variability often collect a limited number of measurements from each recruited individual (e.g., <5). With such low numbers, statistical tests for normality are likely to falsely accept the hypothesis that measurements are normally distributed [107], allowing researchers to carry on the evaluation of a

test while they should have stopped (as current guidelines recommend). Outliers should not be removed from the data prior to analysis, as it is known that they often occur in real life, and may indicate difficulties when performing a measurement. Therefore, the removal of outliers may lead to underestimating the "true" measurement error of a test, and in turn to false conclusions regarding whether a test is fit for use in practice.

With respect to the statistical analysis of primary studies, the thesis introduced alternative methods for providing less biased variability estimates compared to the standard normality-based methods, for measurements expressed as counts rather than continuous. These methods will allow researchers to draw more robust conclusions on the reliability and measurement error of a count-based test, and in turn whether the test is fit for use in clinical practice.

Furthermore, the thesis indicated that systematic reviews examining the reliability and measurement error of tests were not conducted and reported at the desired level. The findings of this review need to be communicated to researchers in this field. Researchers need to understand the importance of using appropriate methods for conducting and reporting such reviews, as such methods ensure that potential bias in the review process is minimised, and in turn allow decisions on reliability and measurement error to be made based on high quality and well reported evidence.

The thesis also identified important flaws regarding how estimates of reliability and measurement error reported across primary studies are combined, and allowed better estimation of a weighted average estimate of reliability and measurement error, by providing meta-analytic methods for key parameters used in the field. These methods will not just produce less biased and more precise summary estimates of reliability and measurement error, enabling researchers to provide more robust summary evidence based on a whole body of research, but will also increase the number of reviews performing a meta-analysis of estimates reported in the primary studies, which was found to be very low.

Finally, the thesis was the first to provide clear evidence on how the reproducibility of the focus score changes for a larger number of glands within the biopsy. The findings presented in Chapter 9 will help clinicians decide on what the minimum number of biopsied glands should be, so that they make best use of the focus score in future research studies of Sjogren's syndrome, as well as in clinical practice.

10.4. Further work

Current guidelines for primary studies examining the variability of laboratory tests need updating, particularly with respect to the target population, the assessment of normality, and the approach taken for outliers. It is also essential that guidelines for physiologic and imaging tests are developed, so that researchers are clear about the steps they need to follow when considering such tests. Furthermore, new methods for estimating the number of individuals required for a primary study are needed, additionally covering the case were inter/intra-observer variability is expected to be present, and accounting for a trade-off between a high reliability and low measurement error.

The findings of the methodological review (Chapter 4) emphasize the need for more specific guidance, both for conducting and reporting systematic reviews examining the reliability and measurement error of biomarkers, as well as appraising the methodological quality of the primary studies.

The new statistical methods for primary analysis and meta-analysis that were developed and presented in the thesis performed well when evaluated in case-study analyses, but require further evaluation by simulation. The extent to which methods proposed for the meta-analysis of the standard error of measurement and the coefficient of variation are affected by different parameter inputs, different numbers of included studies, or different numbers of individuals recruited within studies should be evaluated across different simulated scenarios. For methods proposed for the primary analysis of count measurements, simulation work should compare the count-based models

to the standard linear regression model across a variety of scenarios (e.g., different input values for the number of individuals, measurements per individual, mean and variability at different levels), so that recommendations for the use of count-based models over linear models are provided.

Finally, further development of methods for meta-analysis are also needed, specifically with respect to calculating a 95% prediction interval for key parameters of measurement error, such as the coefficient of variation and the limits of agreement; and combining estimates of the reliability and measurement error of count-based tests (as methods presented in the thesis are restricted to continuous measurements only).

10.5. Conclusions

Knowledge on the reliability and measurement error of biomarkers is required, in order to inform decisions about whether biomarkers are fit for use in medical research and practice. It is essential that primary studies and systematic reviews examining the reliability and measurement error of biomarkers are designed and analysed at a high methodological standard. This thesis provided an initiative for improving how primary studies are designed, and how systematic reviews are conducted and reported, by highlighting flaws and gaps in current practice. Furthermore, statistical methods were presented and evaluated for two under-researched areas in the field: the meta-analysis of estimates expressing the reliability and measurement error of continuous biomarkers; and the primary analysis of studies examining the reliability and measurement error of biomarkers expressed as counts rather than continuous. The methods proposed for primary analysis aim to provide researchers less biased estimates of the reliability and measurement error of a count-based test, compared to the current methods used, helping them decide whether the test is fit for use in medical research and practice. The methods for meta-analysis aim to provide robust summary evidence of the reliability and measurement error of continuous-based tests, allowing conclusions regarding their use in practice to be based on a whole body of research. Although further work is still

required, the thesis is a major step towards highlighting the need for better research studies, as well as expanding the current statistical methodology used in the field.

References

1. Strimbu, K. and J.A. Tavel, *What are biomarkers?* Curr Opin HIV AIDS, 2010. **5**(6): p. 463-6.
2. Califf, R.M., *Biomarker definitions and their applications.* Experimental biology and medicine (Maywood, N.J.), 2018. **243**(3): p. 213-221.
3. *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.* Clin Pharmacol Ther, 2001. **69**(3): p. 89-95.
4. World Health, O. and S. International Programme on Chemical, *Biomarkers and risk assessment : concepts and principles / published under the joint sponsorship of the United Nations environment Programme, the International Labour Organisation, and the World Health Organization*. 1993, World Health Organization: Geneva.
5. FDA-NIH Biomarker Working Group. BEST (Biomarkers, E., and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US); 2016-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK326791/ Co-published by National Institutes of Health (US), Bethesda (MD).
6. Siu, A.L., *Screening for high blood pressure in adults: U.S. Preventive Services Task Force recommendation statement.* Ann Intern Med, 2015. **163**(10): p. 778-86.
7. Benton, S.C., H.E. Seaman, and S.P. Halloran, *Faecal occult blood testing for colorectal cancer screening: the past or the future.* Curr Gastroenterol Rep, 2015. **17**(2): p. 428.
8. Roberts, S.G., et al., *PSA doubling time as a predictor of clinical progression after biochemical failure following radical prostatectomy for prostate cancer.* Mayo Clin Proc, 2001. **76**(6): p. 576-81.
9. Gundogdu, F., et al., *The role of serum CA-125 levels and CA-125 tissue expression positivity in the prediction of the recurrence of stage III and IV epithelial ovarian tumors (CA-125 levels and tissue CA-125 in ovarian tumors).* Arch Gynecol Obstet, 2011. **283**(6): p. 1397-402.
10. Hudson, S. and S. Pettit, *What is 'normal' left ventricular ejection fraction?* Heart, 2020. **106**(18): p. 1445-1446.
11. Roush, G.C. and D.A. Sica, *Diuretics for Hypertension: A Review and Update.* Am J Hypertens, 2016. **29**(10): p. 1130-7.
12. *Revascularization versus Medical Therapy for Renal-Artery Stenosis.* New England Journal of Medicine, 2009. **361**(20): p. 1953-1962.
13. Bartlett, J.W. and C. Frost, *Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables.* Ultrasound Obstet Gynecol, 2008. **31**(4): p. 466-75.
14. Harvill, L.M., *Standard Error of Measurement.* Educational Measurement: Issues and Practice, 1991. **10**: p. 33-41.
15. de Vet, H.C.W., et al., *Measurement in Medicine: A Practical Guide*. Practical Guides to Biostatistics and Epidemiology. 2011, Cambridge: Cambridge University Press.
16. Atkinson, G. and A.M. Nevill, *Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine.* Sports Med, 1998. **26**(4): p. 217-38.
17. Fraser, C.G., *Biological Variation: From Principles to Practice*. 2001.
18. Conroy, R.M., et al., *Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project.* Eur Heart J, 2003. **24**(11): p. 987-1003.
19. Hippisley-Cox, J., et al., *Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study.* Bmj, 2007. **335**(7611): p. 136.
20. Badeli, H. and F. Assadi, *Strategies to reduce pitfalls in measuring blood pressure.* Int J Prev Med, 2014. **5**(Suppl 1): p. S17-20.
21. Gao, T., C. Demino, and J.R. Fowler, *Ultrasound Measurement Error and Its Implications for Carpal Tunnel Syndrome Diagnosis.* Hand (N Y), 2022. **17**(4): p. 635-638.
22. Butler, J., S.D. Anker, and M. Packer, *Redefining Heart Failure With a Reduced Ejection Fraction.* Jama, 2019. **322**(18): p. 1761-1762.

23. Triposkiadis, F., et al., *The continuous heart failure spectrum: moving beyond an ejection fraction classification.* Eur Heart J, 2019. **40**(26): p. 2155-2163.
24. Yancy, C.W., et al., *2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines.* J Am Coll Cardiol, 2013. **62**(16): p. e147-239.
25. Ponikowski, P., et al., *2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC.* Eur Heart J, 2016. **37**(27): p. 2129-2200.
26. Fraser, C.G. and E.K. Harris, *Generation and application of data on biological variation in clinical chemistry.* Crit Rev Clin Lab Sci, 1989. **27**(5): p. 409-37.
27. Braga, F. and M. Panteghini, *Generation of data on within-subject biological variation in laboratory medicine: An update.* Critical Reviews in Clinical Laboratory Sciences, 2016. **53**(5): p. 313-325.
28. Tagmouti, S., et al., *Reproducibility of interferon gamma (IFN-γ) release Assays. A systematic review.* Ann Am Thorac Soc, 2014. **11**(8): p. 1267-76.
29. Weiner, O.M. and J.J. McGrath, *Test-Retest Reliability of Pediatric Heart Rate Variability: A Meta-Analysis.* J Psychophysiol, 2017. **31**(1): p. 6-28.
30. Yoon, S.H., et al., *Observer variability in RECIST-based tumour burden measurements: a meta-analysis.* Eur J Cancer, 2016. **53**: p. 5-15.
31. *How a Meta-Analysis Works*, in *Introduction to Meta-Analysis*. 2009. p. 1-7.
32. Young, D.S., E.K. Harris, and E. Cotlove, *Biological and analytic components of variation in long-term studies of serum constituents in normal subjects. IV. Results of a study designed to eliminate long-term analytic deviations.* Clin Chem, 1971. **17**(5): p. 403-10.
33. Savva, C., et al., *Test-retest reliability of handgrip strength measurement using a hydraulic hand dynamometer in patients with cervical radiculopathy.* J Manipulative Physiol Ther, 2014. **37**(3): p. 206-10.
34. Johenning, A.R., T.G. Karrison, and W.M. Barron, *Interobserver Variability in the Measurement of Diastolic Blood Pressure in Pregnancy.* Hypertension in Pregnancy, 1995. **14**(3): p. 301-311.
35. Preisner, F., et al., *Reliability and reproducibility of sciatic nerve magnetization transfer imaging and T2 relaxometry.* Eur Radiol, 2021. **31**(12): p. 9120-9130.
36. Kottner, J., et al., *Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed.* J Clin Epidemiol, 2011. **64**(1): p. 96-106.
37. Mokkink, L.B., 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 Medical Research Methodology, 2020. **20**(1): p. 293.
38. Banfi, G. and G. Lippi, *The impact of preanalytical variability in clinical trials: are we underestimating the issue?* Annals of Translational Medicine, 2016. **4**(3): p. 59.
39. Lippi, G., et al., *Preanalytical variability: the dark side of the moon in laboratory testing.* Clinical Chemistry and Laboratory Medicine (CCLM), 2006. **44**(4): p. 358-365.
40. Feeney, T. and C. Poole, *Self-testing for covid-19.* BMJ, 2022. **378**: p. o2055.
41. Fraser, C.G., *Reference change values.* Clin Chem Lab Med, 2011. **50**(5): p. 807-12.
42. Fraser, C.G., *Improved Monitoring of Differences in Serial Laboratory Results.* Clinical Chemistry, 2011. **57**(12): p. 1635-1637.
43. Giraudeau, B. and J.Y. Mary, *Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient.* Stat Med, 2001. **20**(21): p. 3205-14.
44. Cochran, W., *The distribution of the largest of a set of estimated variances as a fraction of their total.* Annals of Human Genetics, 2011. **11**: p. 47-52.

45. Dixon, W.J., *Processing Data for Outliers.* Biometrics, 1953. **9**(1): p. 74-89.
46. Shapiro, S.S. and M.B. Wilk, *An Analysis of Variance Test for Normality (Complete Samples).* Biometrika, 1965. **52**: p. 591-611.
47. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. **1**(8476): p. 307-10.
48. Thompson, S.K., *Sampling*. 1992: New York: Wiley.
49. Shrout, P.E. and J.L. Fleiss, *Intraclass correlations: uses in assessing rater reliability.* Psychol Bull, 1979. **86**(2): p. 420-8.
50. Graybill, F.A. and C.-M. Wang, *Confidence Intervals on Nonnegative Linear Combinations of Variances.* Journal of the American Statistical Association, 1980. **75**(372): p. 869-873.
51. Vaz, S., et al., *The case for using the repeatability coefficient when calculating test-retest reliability.* PLoS One, 2013. **8**(9): p. e73990.
52. Cicchetti, D.V., *Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.* Psychological Assessment, 1994. **6**(4): p. 284-290.
53. Koo, T.K. and M.Y. Li, *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.* Journal of chiropractic medicine, 2016. **15**(2): p. 155-163.
54. Martin, J., et al., *Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care.* Trials, 2016. **17**: p. 402.
55. McGraw, K.O. and S.P. Wong, *Forming inferences about some intraclass correlation coefficients.* Psychological Methods, 1996. **1**(1): p. 30-46.
56. Euser, A.M., F.W. Dekker, and S. le Cessie, *A practical approach to Bland-Altman plots and variation coefficients for log transformed variables.* J Clin Epidemiol, 2008. **61**(10): p. 978-82.
57. Pedersen, J. and J. Liu, *Child Mortality Estimation: Appropriate Time Periods for Child Mortality Estimates from Full Birth Histories.* PLOS Medicine, 2012. **9**(8): p. e1001289.
58. Aronhime, S., et al., *DCE-MRI of the liver: effect of linear and nonlinear conversions on hepatic perfusion quantification and reproducibility.* J Magn Reson Imaging, 2014. **40**(1): p. 90-8.
59. McKay, A.T., *Distribution of the Coefficient of Variation and the Extended "t" Distribution.* Journal of the Royal Statistical Society, 1932. **95**(4): p. 695-698.
60. Bland, J.M. and D.G. Altman, *Statistics Notes: Measurement error proportional to the mean.* BMJ, 1996. **313**(7049): p. 106.
61. Cole, T.J., *Sympercents: symmetric percentage differences on the 100 log(e) scale simplify the presentation of log transformed data.* Stat Med, 2000. **19**(22): p. 3109-25.
62. Harris, E.K., *Statistical aspects of reference values in clinical pathology.* Prog Clin Pathol, 1981. **8**: p. 45-66.
63. Simundic, A.-M., et al., *Terms and Symbols Used in Studies on Biological Variation: The Need for Harmonization.* Clinical Chemistry, 2015. **61**(2): p. 438-439.
64. Harris, E.K. and T. Yasaka, *On the calculation of a "reference change" for comparing two consecutive measurements.* Clin Chem, 1983. **29**(1): p. 25-30.
65. Fokkema, M.R., et al., *Reference change values for brain natriuretic peptides revisited.* Clin Chem, 2006. **52**(8): p. 1602-3.
66. Altman, D.G. and J.M. Bland, *Measurement in Medicine: The Analysis of Method Comparison Studies.* Journal of the Royal Statistical Society. Series D (The Statistician), 1983. **32**(3): p. 307-317.
67. Fisher, R.A., *Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population.* Biometrika, 1915. **10**(4): p. 507-521.
68. Schober, P., C. Boer, and L.A. Schwarte, *Correlation Coefficients: Appropriate Use and Interpretation.* Anesthesia & Analgesia, 2018. **126**(5): p. 1763-1768.

69. Mukaka, M.M., *Statistics corner: A guide to appropriate use of correlation coefficient in medical research.* Malawi medical journal : the journal of Medical Association of Malawi, 2012. **24**(3): p. 69-71.
70. Cohen, J., *A Coefficient of Agreement for Nominal Scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 37-46.
71. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.
72. Fleiss, J.L., *Statistical Methods for Rates and Proportions. Second Edition*. 1981: Wiley, John and Sons, Incorporated, New York, N.Y.
73. Cicchetti, D.V. and J.L. Fleiss, *Comparison of the Null Distributions of Weighted Kappa and the C Ordinal Statistic.* Applied Psychological Measurement, 1977. **1**(2): p. 195-201.
74. Røraas, T., P.H. Petersen, and S. Sandberg, *Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals.* Clin Chem, 2012. **58**(9): p. 1306-13.
75. Doob, J.L., *The Limiting Distributions of Certain Statistics.* The Annals of Mathematical Statistics, 1935. **6**(3): p. 160-169, 10.
76. Curto, J.D. and J.C. Pinto, *The coefficient of variation asymptotic distribution in the case of non-iid random variables.* Journal of Applied Statistics, 2009. **36**: p. 21 - 32.
77. Bobko, P. and A. Rieck, *Large sample estimators for standard errors of functions of correlation coefficients.* Applied Psychological Measurement, 1980. **4**(3): p. 385-398.
78. Efron, B., *Bootstrap Methods: Another Look at the Jackknife.* The Annals of Statistics, 1979. **7**(1): p. 1-26, 26.
79. Campbell, M.K. and D.J. Torgerson, *Bootstrapping: estimating confidence intervals for cost-effectiveness ratios.* QJM: An International Journal of Medicine, 1999. **92**(3): p. 177-182.
80. Poi, B.P., *From the help desk: Some bootstrapping techniques.* Stata Journal, 2004. **4**(3): p. 312-328.
81. Bossuyt, P.M., et al., *STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies.* BMJ : British Medical Journal, 2015. **351**: p. h5527.
82. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* BMJ : British Medical Journal, 2015. **350**: p. g7594.
83. Moher, D., et al., *CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials.* BMJ, 2010. **340**: p. c869.
84. Cuschieri, S., *The STROBE guidelines.* Saudi journal of anaesthesia, 2019. **13**(Suppl 1): p. S31-S34.
85. Bartlett, W.A., et al., *A checklist for critical appraisal of studies of biological variation.* Clin Chem Lab Med, 2015. **53**(6): p. 879-85.
86. Riley, R.D., P.C. Lambert, and G. Abo-Zaid, *Meta-analysis of individual participant data: rationale, conduct, and reporting.* BMJ, 2010. **340**: p. c221.
87. Riley, R.D., et al., *A guide to systematic review and meta-analysis of prognostic factor studies.* BMJ, 2019. **364**: p. k4597.
88. Hartzes, A.M. and C.J. Morgan, *Meta-analysis for diagnostic tests.* Journal of Nuclear Cardiology, 2019. **26**(1): p. 68-71.
89. DerSimonian, R. and N. Laird, *Meta-analysis in clinical trials revisited.* Contemp Clin Trials, 2015. **45**(Pt A): p. 139-45.
90. Borenstein, M., et al., *A basic introduction to fixed-effect and random-effects models for meta-analysis.* Res Synth Methods, 2010. **1**(2): p. 97-111.
91. *Fixed-Effect Model*, in *Introduction to Meta-Analysis*. 2009. p. 63-67.
92. *Random-Effects Model*, in *Introduction to Meta-Analysis*. 2009. p. 69-75.
93. Cochran, W.G., *The comparison of percentages in matched samples.* Biometrika, 1950. **37**(3-4): p. 256-66.

94. Hardy, R.J. and S.G. Thompson, *Detecting and describing heterogeneity in meta-analysis.* Stat Med, 1998. **17**(8): p. 841-56.
95. Dickersin, K. and J.A. Berlin, *Meta-analysis: state-of-the-science.* Epidemiol Rev, 1992. **14**: p. 154-76.
96. Higgins, J.P.T., et al., *Measuring inconsistency in meta-analyses.* BMJ, 2003. **327**(7414): p. 557-560.
97. Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook.
98. Borenstein, M., et al., *Basics of meta-analysis: I(2) is not an absolute measure of heterogeneity.* Res Synth Methods, 2017. **8**(1): p. 5-18.
99. Migliavaca, C.B., et al., *Meta-analysis of prevalence: I(2) statistic and how to deal with heterogeneity.* Res Synth Methods, 2022. **13**(3): p. 363-367.
100. Riley, R.D., J.P.T. Higgins, and J.J. Deeks, *Interpretation of random effects meta-analyses.* BMJ, 2011. **342**: p. d549.
101. Jackson, D., J. Bowden, and R. Baker, *How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts?* Journal of Statistical Planning and Inference, 2010. **140**(4): p. 961-970.
102. Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook.
103. Rücker, G., et al., *Undue reliance on I(2) in assessing heterogeneity may mislead.* BMC Med Res Methodol, 2008. **8**: p. 79.
104. Simmonds, M.C., et al., *Meta-analysis of individual patient data from randomized trials: a review of methods used in practice.* Clin Trials, 2005. **2**(3): p. 209-17.
105. *Kolmogorov–Smirnov Test*, in *The Concise Encyclopedia of Statistics*. 2008, Springer New York: New York, NY. p. 283-287.
106. Feng, C., et al., *Log-transformation and its implications for data analysis.* Shanghai Arch Psychiatry, 2014. **26**(2): p. 105-9.
107. Ghasemi, A. and S. Zahediasl, *Normality tests for statistical analysis: a guide for non-statisticians.* Int J Endocrinol Metab, 2012. **10**(2): p. 486-9.
108. Stokes, M., *Reliability and Repeatability of Methods for Measuring Muscle in Physiotherapy.* Physiotherapy Practice, 1985. **1**(2): p. 71-76.
109. Bohannon, R.W., *Muscle strength: clinical and prognostic value of hand-grip dynamometry.* Curr Opin Clin Nutr Metab Care, 2015. **18**(5): p. 465-70.
110. Lee, S.H. and H.S. Gong, *Measurement and Interpretation of Handgrip Strength for Research on Sarcopenia and Osteoporosis.* Journal of bone metabolism, 2020. **27**(2): p. 85-96.
111. Cruz-Jentoft, A.J., et al., *Sarcopenia: revised European consensus on definition and diagnosis.* Age and Ageing, 2018. **48**(1): p. 16-31.
112. Cruz-Jentoft, A.J. and A.A. Sayer, *Sarcopenia.* The Lancet, 2019. **393**(10191): p. 2636-2646.
113. Clifford, M.S., et al., *Grip strength dynamometry: reliability and validity for adults with upper limb burns.* Burns, 2013. **39**(7): p. 1430-6.
114. Ling, C.H.Y., et al., *Handgrip strength and mortality in the oldest old population: the Leiden 85-plus study.* CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 2010. **182**(5): p. 429-435.
115. Park, S.W., et al., *Decreased muscle strength and quality in older adults with type 2 diabetes: the health, aging, and body composition study.* Diabetes, 2006. **55**(6): p. 1813-8.
116. Cetinus, E., et al., *Hand grip strength in patients with type 2 diabetes mellitus.* Diabetes Res Clin Pract, 2005. **70**(3): p. 278-86.

117. Beenakker, K.G., et al., *Patterns of muscle strength loss with age in the general population and patients with a chronic inflammatory state.* Ageing Res Rev, 2010. **9**(4): p. 431-6.
118. Bohannon, R.W., *Adequacy of hand-grip dynamometry for characterizing upper limb strength after stroke.* Isokinetics and Exercise Science, 2004. **12**: p. 263-265.
119. Soyupek, F., et al., *Androgen deprivation therapy for prostate cancer: effects on hand function.* Urol Oncol, 2008. **26**(2): p. 141-6.
120. Cortopassi, F., et al., *Resting handgrip force and impaired cardiac function at rest and during exercise in COPD patients.* Respir Med, 2011. **105**(5): p. 748-54.
121. Yorke, A.M., et al., *The impact of multimorbidity on grip strength in adults age 50 and older: Data from the health and retirement survey (HRS).* Arch Gerontol Geriatr, 2017. **72**: p. 164-168.
122. Fess, E. and C. Moran, *American Society of Hand Therapists Clinical Assessment Recommendations*. 1981.
123. American Society of Hand, T., *Clinical assessment recommendations*. 1992, Chicago (401 N. Michigan Ave., Chicago IL 60611-4267): The Society.
124. Roberts, H.C., et al., *A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach.* Age and Ageing, 2011. **40**(4): p. 423-429.
125. Efron B, T.R., *An introduction to the bootstrap*, ed. C.H. New York. 1993.
126. Mooney, C.Z., *Bootstrapping : a nonparametric approach to statstical inference*, R. Duvall, Editor. 1993, Sage Publications: Newbury Park, Calif. :.
127. Streiner, D.L., G.R. Norman, and J. Cairney, *Health Measurement ScalesA practical guide to their development and use: A practical guide to their development and use*. 2015, Oxford, UK: Oxford University Press.
128. Savva, C., et al., *Test-Retest Reliability of Handgrip Strength as an Outcome Measure in Patients With Symptoms of Shoulder Impingement Syndrome.* J Manipulative Physiol Ther, 2018. **41**(3): p. 252-257.
129. Alfonso-Rosa, R.M., et al., *Test-retest reliability and minimal detectable change scores for fitness assessment in older adults with type 2 diabetes.* Rehabil Nurs, 2014. **39**(5): p. 260-8.
130. Oxman, A.D. and G.H. Guyatt, *The science of reviewing research.* Ann N Y Acad Sci, 1993. **703**: p. 125-33; discussion 133-4.
131. Antman, E.M., et al., *A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction.* Jama, 1992. **268**(2): p. 240-8.
132. Leeflang, M.M., et al., *Systematic reviews of diagnostic test accuracy.* Ann Intern Med, 2008. **149**(12): p. 889-97.
133. Mokkink, L.B., et al., *Evaluation of the methodological quality of systematic reviews of health status measurement instruments.* Qual Life Res, 2009. **18**(3): p. 313-33.
134. Mokkink, L.B., et al., *The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study.* Qual Life Res, 2010. **19**(4): p. 539-49.
135. Terwee, C.B., et al., *Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments.* Qual Life Res, 2009. **18**(8): p. 1115-23.
136. Lucas, N., et al., *The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL).* BMC Med Res Methodol, 2013. **13**: p. 111.
137. de Langen, A.J., et al., *Repeatability of 18F-FDG uptake measurements in tumors: a metaanalysis.* J Nucl Med, 2012. **53**(5): p. 701-8.
138. Kramer, G.M., et al., *Repeatability of quantitative (18)F-FLT uptake measurements in solid tumors: an individual patient data multi-center meta-analysis.* Eur J Nucl Med Mol Imaging, 2018. **45**(6): p. 951-961.

139. Powden, C.J., J.M. Hoch, and M.C. Hoch, *Reliability and minimal detectable change of the weight-bearing lunge test: A systematic review.* Man Ther, 2015. **20**(4): p. 524-32.
140. Rondoni, A., et al., *Intrarater and Inter-rater Reliability of Active Cervical Range of Motion in Patients With Nonspecific Neck Pain Measured With Technological and Common Use Devices: A Systematic Review With Meta-regression.* J Manipulative Physiol Ther, 2017. **40**(8): p. 597-608.
141. Welton, T., et al., *Reproducibility of graph-theoretic brain network metrics: a systematic review.* Brain connectivity, 2015. **5**(4): p. 193-202.
142. Hedges LV, O.I., *Statistical methods for meta-analysis*. 1985, San Diego: Academic Press.
143. Schmidt, F., Hunter, J., *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 2015: 55 City Road, London.
144. Rabelo, M., et al., *Reliability of muscle strength assessment in chronic post-stroke hemiparesis: a systematic review and meta-analysis.* Top Stroke Rehabil, 2016. **23**(1): p. 26-36.
145. Williamson, P.R., et al., *Meta-analysis of method comparison studies.* Stat Med, 2002. **21**(14): p. 2013-25.
146. Sun, S., *Meta-analysis of Cohen's kappa.* Health Services and Outcomes Research Methodology, 2011. **11**(3): p. 145-163.
147. Lange, T., et al., *Reliability of specific physical examination tests for the diagnosis of shoulder pathologies: a systematic review and meta-analysis.* Br J Sports Med, 2017. **51**(6): p. 511-518.
148. Everitt, B.S., *Moments of the statistics kappa and weighted kappa*. Vol. 21(1). 1968: British Journal of Mathematical and Statistical Psychology. 97-103.
149. Reavis, K.M., et al., *Meta-Analysis of Distortion Product Otoacoustic Emission Retest Variability for Serial Monitoring of Cochlear Function in Adults.* Ear and hearing, 2015. **36**(5): p. e251-e260.
150. Kristof, W., *The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts.* Psychometrika, 1963. **28**(3): p. 221-238.
151. Rozema, J.J., et al., *Overview of the repeatability, reproducibility, and agreement of the biometry values provided by various ophthalmic devices.* American journal of ophthalmology, 2014. **158**(6): p. 1111-1120.e1.
152. Serai, S.D., et al., *Repeatability of MR Elastography of Liver: A Meta-Analysis.* Radiology, 2017. **285**(1): p. 92-100.
153. Aarsand, A.K., et al., *The biological variation data critical appraisal checklist: A standard for evaluating studies on biological variation.* Clinical Chemistry, 2018. **64**(3): p. 501-514.
154. Cavaleri, R., S.M. Schabrun, and L.S. Chipchase, *The number of stimuli required to reliably assess corticomotor excitability and primary motor cortical representations using transcranial magnetic stimulation (TMS): a systematic review and meta-analysis.* Syst Rev, 2017. **6**(1): p. 48.
155. Chamorro, C., et al., *Absolute Reliability and Concurrent Validity of Hand Held Dynamometry and Isokinetic Dynamometry in the Hip, Knee and Ankle Joint: Systematic Review and Meta-analysis.* Open Med (Wars), 2017. **12**: p. 359-375.
156. González-Lao, E., et al., *Systematic review of the biological variation data for diabetes related analytes.* Clin Chim Acta, 2019. **488**: p. 61-67.
157. Hunter, D.J., et al., *Responsiveness and reliability of MRI in knee osteoarthritis: a meta-analysis of published evidence.* Osteoarthritis Cartilage, 2011. **19**(5): p. 589-605.
158. Kleijn, S.A., et al., *A meta-analysis of left ventricular dyssynchrony assessment and prediction of response to cardiac resynchronization therapy by three-dimensional echocardiography.* Eur Heart J Cardiovasc Imaging, 2012. **13**(9): p. 763-75.
159. Navarro, I., B.N.D. Rosa, and C.T. Candotti, *Anatomical reference marks, evaluation parameters and reproducibility of surface topography for evaluating the adolescent*

idiopathic scoliosis: a systematic review with meta-analysis. Gait Posture, 2019. **69**: p. 112-120.
160. Reichmann, W.M., et al., *Responsiveness to change and reliability of measurement of radiographic joint space width in osteoarthritis of the knee: a systematic review.* Osteoarthritis Cartilage, 2011. **19**(5): p. 550-6.
161. Salamh, P.A., et al., *The reliability, validity, and methodologic quality of measurements used to quantify posterior shoulder tightness: a systematic review of the literature with meta-analysis.* J Shoulder Elbow Surg, 2019. **28**(1): p. 178-185.
162. van Tulder, M., et al., *Updated method guidelines for systematic reviews in the cochrane collaboration back review group.* Spine (Phila Pa 1976), 2003. **28**(12): p. 1290-9.
163. Lemeshow, A.R., et al., *Searching one or two databases was insufficient for meta-analysis of observational studies.* J Clin Epidemiol, 2005. **58**(9): p. 867-73.
164. Waffenschmidt, S., et al., *Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review.* BMC Medical Research Methodology, 2019. **19**(1): p. 132.
165. Shea, B.J., et al., *AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both.* BMJ, 2017. **358**: p. j4008.
166. Tipton, E. and J. Shuster, *A framework for the meta-analysis of Bland-Altman studies based on a limits of agreement approach.* Statistics in medicine, 2017. **36**(23): p. 3621-3635.
167. Fienberg, S.E., Biometrics, 1971. **27**(1): p. 238-241.
168. OpenStax College, Introductory Statistics. OpenStax College. 19 September 2013
169. Crawshaw, J.C.J., *A concise course in advanced level statistics : with worked examples*. 2001, Nelson Thornes.
170. *Effect Sizes Based on Binary Data (2×2 Tables)*, in *Introduction to Meta-Analysis*. 2009. p. 33-39.
171. Terwee, C.B., et al., *Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments.* Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 2009. **18**(8): p. 1115-1123.
172. McInnes, M.D.F., et al., *Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement.* Jama, 2018. **319**(4): p. 388-396.
173. Bechtol, C.O., *Grip test; the use of a dynamometer with adjustable handle spacings.* J Bone Joint Surg Am, 1954. **36-a**(4): p. 820-4; passim.
174. Jarit, P., *Dominant-hand to nondominant-hand grip-strength ratios of college baseball players.* Journal of Hand Therapy, 1991. **4**(3): p. 123-126.
175. Lunde, B.K., W.D. Brewer, and P.A. Garcia, *Grip strength of college women.* Arch Phys Med Rehabil, 1972. **53**(10): p. 491-3.
176. Schmidt, R.T. and J.V. Toews, *Grip strength as measured by the Jamar dynamometer.* Arch Phys Med Rehabil, 1970. **51**(6): p. 321-7.
177. Mathiowetz, V., C. Rennells, and L. Donahoe, *Effect of elbow position on grip and key pinch strength.* J Hand Surg Am, 1985. **10**(5): p. 694-7.
178. Richards, L. and P. Palmiter-Thomas, *Grip Strength Measurement: A Critical Review of Tools, Methods, and Clinical Utility.* 1996. **8**(1-2): p. 87-109.
179. Shechtman, O., et al., *Reliability and validity of the BTE-Primus grip tool.* J Hand Ther, 2003. **16**(1): p. 36-42.
180. Mathiowetz, V., et al., *Reliability and validity of grip and pinch strength evaluations.* J Hand Surg Am, 1984. **9**(2): p. 222-6.
181. Hamilton, A., R. Balnave, and R. Adams, *Grip strength testing reliability.* J Hand Ther, 1994. **7**(3): p. 163-70.

182. Coldham, F., J. Lewis, and H. Lee, *The reliability of one vs. three grip trials in symptomatic and asymptomatic subjects.* J Hand Ther, 2006. **19**(3): p. 318-26; quiz 327.
183. Alencar, M.A., et al., *Handgrip strength in elderly with dementia: study of reliability.* Rev Bras Fisioter, 2012. **16**(6): p. 510-4.
184. Reuter, S.E., N. Massy-Westropp, and A.M. Evans, *Reliability and validity of indices of hand-grip strength and endurance.* Aust Occup Ther J, 2011. **58**(2): p. 82-7.
185. Schaubert, K.L. and R.W. Bohannon, *Reliability and validity of three strength measures obtained from community-dwelling elderly persons.* J Strength Cond Res, 2005. **19**(3): p. 717-20.
186. Dunn, J.M., *Reliability of selected psychomotor measures with mentally retarded adult males.* Percept Mot Skills, 1978. **46**(1): p. 295-301.
187. Reddon, J.R., et al., *Hand dynamometer: effects of trials and sessions.* Percept Mot Skills, 1985. **61**(3 Pt 2): p. 1195-8.
188. Hamilton, G.F., C. McDonald, and T.C. Chenier, *Measurement of grip strength: validity and reliability of the sphygmomanometer and jamar grip dynamometer.* J Orthop Sports Phys Ther, 1992. **16**(5): p. 215-9.
189. Huang, S.L., et al., *Optimal scoring methods of hand-strength tests in patients with stroke.* Int J Rehabil Res, 2011. **34**(2): p. 178-80.
190. Trutschnigg, B., et al., *Precision and reliability of strength (Jamar vs. Biodex handgrip) and body composition (dual-energy X-ray absorptiometry vs. bioimpedance analysis) measurements in advanced cancer patients.* Appl Physiol Nutr Metab, 2008. **33**(6): p. 1232-9.
191. Barden, H.L., et al., *Test-retest reliability of computerised hand dynamometry in adults with acquired brain injury.* Aust Occup Ther J, 2012. **59**(4): p. 319-27.
192. Legendre, P., *Species Associations: The Kendall Coefficient of Concordance Revisited.* Journal of agricultural, biological, and environmental statistics, 2005. **10**(2): p. 226-245.
193. Nitschke, J.E., et al., *When is a change a genuine change?: A clinically meaningful interpretation of grip strength measurements in healthy and disabled women.* Journal of Hand Therapy, 1999. **12**(1): p. 25-30.
194. Peolsson, A., R. Hedlund, and B. Oberg, *Intra- and inter-tester reliability and reference values for hand strength.* J Rehabil Med, 2001. **33**(1): p. 36-41.
195. Reijnierse, E.M., et al., *Assessment of maximal handgrip strength: how many attempts are needed?* Journal of cachexia, sarcopenia and muscle, 2017. **8**(3): p. 466-474.
196. Guerra, R.S., et al., *Comparison of Jamar and Bodygrip Dynamometers for Handgrip Strength Measurement.* J Strength Cond Res, 2017. **31**(7): p. 1931-1940.
197. Day, S.J. and D.G. Altman, *Blinding in clinical trials and other studies.* BMJ, 2000. **321**(7259): p. 504.
198. Mathiowetz, V., *Effects of three trials on grip and pinch strength measurements.* Journal of Hand Therapy, 1990. **3**(4): p. 195-198.
199. Fisher, B.A., et al., *A review of salivary gland histopathology in primary Sjögren's syndrome with a focus on its potential as a clinical trials biomarker.* Annals of the Rheumatic Diseases, 2015. **74**(9): p. 1645-1650.
200. Koo, H.K., et al., *Systemic White Blood Cell Count as a Biomarker Associated with Severity of Chronic Obstructive Lung Disease.* Tuberc Respir Dis (Seoul), 2017. **80**(3): p. 304-310.
201. Green, J.A., *Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression.* Health Psychol Behav Med, 2021. **9**(1): p. 436-455.
202. Leckie G, B.W., Goldstein H, Merlo J, Austin P, *Variance partitioning in multilevel models for count data.* arXiv, 2019.
203. Jolicoeur, P., *The Poisson distribution*, in *Introduction to Biometry*, P. Jolicoeur, Editor. 1999, Springer US: Boston, MA. p. 124-133.

204. Austin, P.C., et al., *Measures of clustering and heterogeneity in multilevel Poisson regression analyses of rates/count data.* Stat Med, 2018. **37**(4): p. 572-589.
205. MR, S., *Mathematical Handbook of Formulas and Tables.* New York, NY: McGraw-Hill, 1968.
206. McCulloch, C.E., & Searle, S. R., *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics, ed. S.S.W. Walter A. Shewhart. 2001.
207. W.S.), S.G., *Biometrika*. Vol. 5. 1907: Biometrika Trust.
208. Morrell, C.H., *Likelihood Ratio Testing of Variance Components in the Linear Mixed-Effects Model Using Restricted Maximum Likelihood.* Biometrics, 1998. **54**(4): p. 1560-1568.
209. Machowicz, A., et al., *Mediterranean diet and risk of Sjögren's syndrome.* Clin Exp Rheumatol, 2020. **38 Suppl 126**(4): p. 216-221.
210. Fisher, B.A., et al., *Standardisation of labial salivary gland histopathology in clinical trials in primary Sjogren's syndrome.* Ann Rheum Dis, 2017. **76**(7): p. 1161-1168.
211. Juarez, M., et al., *A phase 2 randomized, double-blind, placebo-controlled, proof-of-concept study of oral seletalisib in primary Sjögren's syndrome.* Rheumatology (Oxford), 2021. **60**(3): p. 1364-1375.
212. H., S., *On knowledge of the keratoconjunctivitis sicca. VII. The sicca syndrome--an autoimmune disease.* Acta Ophthalmol (Copenh). **46**: p. 201-206.
213. Murube, J., *Primary and secondary Sjogren-Jones syndromes-historical evolution.* Ocul Surf, 2011. **9**(1): p. 13-6.
214. Tzioufas, A.G. and M. Voulgarelis, *Update on Sjogren's syndrome autoimmune epithelitis: from classification to increased neoplasias.* Best Pract Res Clin Rheumatol, 2007. **21**(6): p. 989-1010.
215. Kroese, F.G.M., E.A. Haacke, and M. Bombardieri, *The role of salivary gland histopathology in primary Sjogren's syndrome: promises and pitfalls.* Clin Exp Rheumatol, 2018. **36 Suppl 112**(3): p. 222-233.
216. Hamaker, E., et al., *Model selection based on information criteria in multilevel modeling*. 2011.
217. Daniels, T.E., et al., *Associations between salivary gland histopathologic diagnoses and phenotypic features of Sjogren's syndrome among 1,726 registry participants.* Arthritis Rheum, 2011. **63**(7): p. 2021-30.
218. Caporali, R., et al., *Safety and usefulness of minor salivary gland biopsy: retrospective analysis of 502 procedures performed at a single center.* Arthritis Rheum, 2008. **59**(5): p. 714-20.
219. de Vet, H.C., et al., *When to use agreement versus reliability measures.* J Clin Epidemiol, 2006. **59**(10): p. 1033-9.
220. Morris, T., I. White, and M. Crowther, *Using simulation studies to evaluate statistical methods.* Statistics in Medicine, 2019. **38**.
221. Devauchelle-Pensec, V., et al., *Which and How Many Patients Should Be Included in Randomised Controlled Trials to Demonstrate the Efficacy of Biologics in Primary Sjogren's Syndrome?* PLoS One, 2015. **10**(9): p. e0133907.
222. Shih, K.C., et al., *Systematic review of randomized controlled trials in the treatment of dry eye disease in Sjogren syndrome.* Journal of inflammation (London, England), 2017. **14**: p. 26-26.
223. Abou El Hassan, M., et al., *Diurnal rhythm in clinical chemistry: An underrated source of variation.* Critical Reviews in Clinical Laboratory Sciences, 2018. **55**(8): p. 516-534.
224. Abou, L., et al., *Clinical Instruments for Measuring Unsupported Sitting Balance in Subjects with Spinal Cord Injury: A Systematic Review.* Topics in spinal cord injury rehabilitation, 2018. **24**(2): p. 177-193.
225. Adhia, D.B., et al., *Validity and reliability of palpation-digitization for non-invasive kinematic measurement - a systematic review.* Manual therapy, 2013. **18**(1): p. 26-34.

226. Ager, A.L., et al., *Shoulder proprioception: How is it measured and is it reliable? A systematic review.* Journal of hand therapy : official journal of the American Society of Hand Therapists, 2017. **30**(2): p. 221-231.
227. Aguilar, H.N., M.C. Battie, and J.L. Jaremko, *MRI-based hip cartilage measures in osteoarthritic and non-osteoarthritic individuals: A systematic review.* RMD Open, 2017. **3**(1): p. e000358.
228. Alcazar, J.L. and M. Jurado, *Three-dimensional ultrasound for assessing women with gynecological cancer: a systematic review.* Gynecologic oncology, 2011. **120**(3): p. 340-6.
229. Aloraini, S.M., et al., *Assessment of spasticity after stroke using clinical measures: a systematic review.* Disability and rehabilitation, 2015. **37**(25): p. 2313-23.
230. Alreni, A.S.E., et al., *Measures of upper limb function for people with neck pain. A systematic review of measurement and practical properties.* Musculoskeletal science & practice, 2017. **29**: p. 155-163.
231. Ammann-Reiffer, C., et al., *Measurement properties of gait-related outcomes in youth with neuromuscular diagnoses: a systematic review.* Physical therapy, 2014. **94**(8): p. 1067-82.
232. Artero, E.G., et al., *Reliability of field-based fitness tests in youth.* International journal of sports medicine, 2011. **32**(3): p. 159-69.
233. Avouac, J., et al., *Validation of the 6 min walk test according to the OMERACT filter: a systematic literature review by the EPOSS-OMERACT group.* Annals of the rheumatic diseases, 2010. **69**(7): p. 1360-3.
234. Balemans, A.C., et al., *Systematic review of the clinimetric properties of laboratory- and field-based aerobic and anaerobic fitness measures in children with cerebral palsy.* Archives of physical medicine and rehabilitation, 2013. **94**(2): p. 287-301.
235. Balzer, J., et al., *Selective voluntary motor control measures of the lower extremity in children with upper motor neuron lesions: a systematic review.* Developmental medicine and child neurology, 2017. **59**(7): p. 699-705.
236. Barrett, E., K. McCreesh, and J. Lewis, *Reliability and validity of non-radiographic methods of thoracic kyphosis measurement: a systematic review.* Manual therapy, 2014. **19**(1): p. 10-7.
237. Bartels, B., J.F. de Groot, and C.B. Terwee, *The six-minute walk test in chronic pediatric conditions: a systematic review of measurement properties.* Physical therapy, 2013. **93**(4): p. 529-41.
238. Basile, F., R. Scionti, and M. Petracca, *Diagnostic reliability of osteopathic tests: A systematic review.* International Journal of Osteopathic Medicine, 2017. **25**: p. 21-29.
239. Beales, L., et al., *Reproducibility of ultrasound measurement of the abdominal aorta.* The British journal of surgery, 2011. **98**(11): p. 1517-25.
240. Beaulieu, L.-D., et al., *Reliability and minimal detectable change of transcranial magnetic stimulation outcomes in healthy adults: A systematic review.* Brain stimulation, 2017. **10**(2): p. 196-213.
241. Bellet, R.N., L. Adams, and N.R. Morris, *The 6-minute walk test in outpatient cardiac rehabilitation: validity, reliability and responsiveness--a systematic review.* Physiotherapy, 2012. **98**(4): p. 277-86.
242. Bennett, H., et al., *Validity of Submaximal Step Tests to Estimate Maximal Oxygen Uptake in Healthy Adults.* Sports medicine (Auckland, N.Z.), 2016. **46**(5): p. 737-50.
243. Berger, A.J., A. Momeni, and A.L. Ladd, *Intra- and interobserver reliability of the Eaton classification for trapeziometacarpal arthritis: a systematic review.* Clinical orthopaedics and related research, 2014. **472**(4): p. 1155-9.
244. Bergquist, R., et al., *Performance-based clinical tests of balance and muscle strength used in young seniors: a systematic literature review.* BMC geriatrics, 2019. **19**(1): p. 9.
245. Bernard, P., et al., *Six minutes walk test for individuals with schizophrenia.* Disability and rehabilitation, 2015. **37**(11): p. 921-7.

246. Bianco, A., et al., *A systematic review to determine reliability and usefulness of the field-based test batteries for the assessment of physical fitness in adolescents - The ASSO Project.* International journal of occupational medicine and environmental health, 2015. **28**(3): p. 445-78.
247. Bieniek, S. and M. Bethge, *The reliability of WorkWell Systems Functional Capacity Evaluation: a systematic review.* BMC musculoskeletal disorders, 2014. **15**: p. 106.
248. Bohannon, R.W., *Test-retest reliability of the five-repetition sit-to-stand test: a systematic review of the literature involving adults.* Journal of strength and conditioning research, 2011. **25**(11): p. 3205-7.
249. Bohannon, R.W., *Test-Retest Reliability of Measurements of Hand-Grip Strength Obtained by Dynamometry from Older Adults: A Systematic Review of Research in the PubMed Database.* The Journal of frailty & aging, 2017. **6**(2): p. 83-87.
250. Bohannon, R.W. and R.H. Crouch, *Two-Minute Step Test of Exercise Capacity: Systematic Review of Procedures, Performance, and Clinimetric Properties.* Journal of geriatric physical therapy (2001), 2019. **42**(2): p. 105-112.
251. Borotikar, B., et al., *Dynamic MRI to quantify musculoskeletal motion: A systematic review of concurrent validity and reliability, and perspectives for evaluation of musculoskeletal disorders.* PloS one, 2017. **12**(12): p. e0189587.
252. Braga, F., et al., *Biological variability of glycated hemoglobin.* Clinica Chimica Acta, 2010. **411**(21-22): p. 1606-1610.
253. Braga, F. and M. Panteghini, *Biologic variability of C-reactive protein: is the available information reliable?* Clinica chimica acta; international journal of clinical chemistry, 2012. **413**(15-16): p. 1179-83.
254. Brink, Y., Q. Louw, and K. Grimmer-Somers, *The quality of evidence of psychometric properties of three-dimensional spinal posture-measuring instruments.* BMC musculoskeletal disorders, 2011. **12**: p. 93.
255. Burgess, F., et al., *Oncology EDGE task force on colorectal cancer outcomes: A systematic review of clinical measures of strength and muscular endurance.* Rehabilitation Oncology, 2016. **34**(1): p. 36-47.
256. Carlsson, H. and E. Rasmussen-Barr, *Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies.* Manual therapy, 2013. **18**(2): p. 103-10.
257. Carobene, A., et al., *A systematic review of data on biological variation for alanine aminotransferase, aspartate aminotransferase and gamma-glutamyl transferase.* Clinical chemistry and laboratory medicine, 2013. **51**(10): p. 1997-2007.
258. Chaabene, H., et al., *Tests for the assessment of sport-specific performance in Olympic combat sports: A systematic review with practical recommendations.* Frontiers in Physiology, 2018. **9**(APR): p. 386.
259. Cheung, P.P., M. Dougados, and L. Gossec, *Reliability of ultrasonography to detect synovitis in rheumatoid arthritis: a systematic literature review of 35 studies (1,415 patients).* Arthritis care & research, 2010. **62**(3): p. 323-34.
260. Childs, J.T., et al., *Methods of determining the size of the adult liver using 2D ultrasound: A systematic review of articles reporting liver measurement techniques.* Journal of Diagnostic Medical Sonography, 2014. **30**(6): p. 296-306.
261. Chiwaridzo, M., et al., *A systematic review investigating measurement properties of physiological tests in rugby.* BMC Sports Science, Medicine and Rehabilitation, 2017. **9**(1): p. 24.
262. Clark, R., et al., *Clinimetric properties of lower limb neurological impairment tests for children and young people with a neurological condition: A systematic review.* PloS one, 2017. **12**(7): p. e0180031.

263. Clark, R.A., et al., *Reliability and validity of the Wii Balance Board for assessment of standing balance: A systematic review.* Gait & posture, 2018. **61**: p. 40-54.
264. Crowley, A.L., et al., *Critical Review of Current Approaches for Echocardiographic Reproducibility and Reliability Assessment in Clinical Research.* Journal of the American Society of Echocardiography : official publication of the American Society of Echocardiography, 2016. **29**(12): p. 1144-1154.e7.
265. Cutolo, M., et al., *Is laser speckle contrast analysis (LASCA) the new kid on the block in systemic sclerosis? A systematic literature review and pilot study to evaluate reliability of LASCA to measure peripheral blood perfusion in scleroderma patients.* Autoimmunity reviews, 2018. **17**(8): p. 775-780.
266. de Albuquerque, P.M.N.M., et al., *Concordance and Reliability of Photogrammetric Protocols for Measuring the Cervical Lordosis Angle: A Systematic Review of the Literature.* Journal of Manipulative and Physiological Therapeutics, 2018. **41**(1): p. 71-80.
267. De Guio, F., et al., *Reproducibility and variability of quantitative magnetic resonance imaging markers in cerebral small vessel disease.* Journal of Cerebral Blood Flow and Metabolism, 2016. **36**(8): p. 1319-1337.
268. de Paula Lima, P.O., et al., *Measurement properties of the pressure biofeedback unit in the evaluation of transversus abdominis muscle activity: a systematic review.* Physiotherapy, 2011. **97**(2): p. 100-6.
269. De Valk, E.J., J.C.A. Noorduyn, and E.L.A.R. Mutsaerts, *How to assess femoral and tibial component rotation after total knee arthroplasty with computed tomography: a systematic review.* Knee surgery, sports traumatology, arthroscopy : official journal of the ESSKA, 2016. **24**(11): p. 3517-3528.
270. Decary, S., et al., *Reliability of physical examination tests for the diagnosis of knee disorders: Evidence from a systematic review.* Manual therapy, 2016. **26**: p. 172-182.
271. Dejong, H.M., et al., *The validity and reliability of using ultrasound elastography to measure cutaneous stiffness, a systematic review.* International Journal of Burns and Trauma, 2017. **7**(7): p. 124-141.
272. Dekkers, K.J.F.M., et al., *Upper extremity strength measurement for children with cerebral palsy: a systematic review of available instruments.* Physical therapy, 2014. **94**(5): p. 609-22.
273. Deng, H. and C.W.P. Li-Tsang, *Measurement of vascularity in the scar: A systematic review.* Burns, 2018.
274. Denteneer, L., et al., *Inter- and Intrarater Reliability of Clinical Tests Associated With Functional Lumbar Segmental Instability and Motor Control Impairment in Patients With Low Back Pain: A Systematic Review.* Archives of physical medicine and rehabilitation, 2017. **98**(1): p. 151-164.e6.
275. Denteneer, L., et al., *Reliability of physical functioning tests in patients with low back pain: a systematic review.* The spine journal : official journal of the North American Spine Society, 2018. **18**(1): p. 190-207.
276. D'Hondt, N.E., et al., *Reliability of Performance-Based Clinical Measurements to Assess Shoulder Girdle Kinematics and Positioning: Systematic Review.* Physical therapy, 2017. **97**(1): p. 124-144.
277. Dobson, F., et al., *Clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments: a systematic review.* Arthritis care & research, 2012. **64**(10): p. 1565-75.
278. Dobson, F., et al., *Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review.* Osteoarthritis and cartilage, 2012. **20**(12): p. 1548-62.
279. Ekpo, E.U. and M.F. McEntee, *Measurement of breast density with digital breast tomosynthesis--a systematic review.* The British journal of radiology, 2014. **87**(1043): p. 20140460.

280. English, C., L. Fisher, and K. Thoirs, *Reliability of real-time ultrasound for measuring skeletal muscle size in human limbs in vivo: a systematic review.* Clinical rehabilitation, 2012. **26**(10): p. 934-44.
281. Fan, A.P., et al., *Comparison of cerebral blood flow measurement with [15 O]-water positron emission tomography and arterial spin labeling magnetic resonance imaging: A systematic review.* Journal of Cerebral Blood Flow and Metabolism, 2015. **36**(5): p. 842-861.
282. Ferreira, J.B., et al., *Accuracy and reproducibility of dental measurements on tomographic digital models: a systematic review and meta-analysis.* Dento maxillo facial radiology, 2017. **46**(7): p. 20160455.
283. Field, D. and R. Livingstone, *Clinical tools that measure sitting posture, seated postural control or functional abilities in children with motor impairments: a systematic review.* Clinical rehabilitation, 2013. **27**(11): p. 994-1004.
284. Fisher, M.I., et al., *Oncology section edge task force on prostate cancer outcomes: A systematic review of clinical measures of strength and muscular endurance.* Rehabilitation Oncology, 2015. **33**(2): p. 37-44.
285. Flamand, V.H., H. Masse-Alarie, and C. Schneider, *Psychometric evidence of spasticity measurement tools in cerebral palsy children and adolescents: a systematic review.* Journal of rehabilitation medicine, 2013. **45**(1): p. 14-23.
286. Fonseca, M.D.C.R., et al., *Functional, motor, and sensory assessment instruments upon nerve repair in adult hands: Systematic review of psychometric properties.* Systematic Reviews, 2018. **7**(1): p. 175.
287. Fotheringham, I., et al., *Comparison of laboratory- and field-based exercise tests for COPD: a systematic review.* International journal of chronic obstructive pulmonary disease, 2015. **10**: p. 625-43.
288. Gebruers, N., et al., *Monitoring of physical activity after stroke: a systematic review of accelerometry-based measures.* Archives of physical medicine and rehabilitation, 2010. **91**(2): p. 288-97.
289. Gengo e Silva, R.C., V.F.A. de Melo, and M.A.M. Lima, *Validity, reliability and accuracy of oscillometric devices, compared with Doppler ultrasound, for determination of the Ankle Brachial Index: An integrative review.* Jornal Vascular Brasileiro, 2014. **13**(1): p. 27-33.
290. Giraud, R., et al., *Reproducibility of transpulmonary thermodilution cardiac output measurements in clinical practice: a systematic review.* Journal of clinical monitoring and computing, 2017. **31**(1): p. 43-51.
291. Golden, S.H., et al., *Reliability of hypothalamic-pituitary-adrenal axis assessment methods for use in population-based studies.* European Journal of Epidemiology, 2011. **26**(7): p. 511-525.
292. Gonzalez-Suarez, C.B., et al., *Inter-rater and intra-rater reliability of sonographic median nerve and wrist measurements.* Journal of Medical Ultrasound, 2018. **26**(1): p. 14-23.
293. Guillaud, A., et al., *Reliability of Diagnosis and Clinical Efficacy of Cranial Osteopathy: A Systematic Review.* PloS one, 2016. **11**(12): p. e0167823.
294. Hafsteinsdottir, T.B., M. Rensink, and M. Schuurmans, *Clinimetric properties of the Timed Up and Go Test for patients with stroke: a systematic review.* Topics in stroke rehabilitation, 2014. **21**(3): p. 197-210.
295. Hernaez, R., et al., *Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: a meta-analysis.* Hepatology (Baltimore, Md.), 2011. **54**(3): p. 1082-1090.
296. Himuro, N., et al., *Easy-to-use clinical measures of walking ability in children and adolescents with cerebral palsy: a systematic review.* Disability and rehabilitation, 2017. **39**(10): p. 957-968.
297. Hoogervorst, P., et al., *Reliability of measurements of the fractured clavicle: a systematic review.* Systematic reviews, 2017. **6**(1): p. 223.

298. Hulteen, R.M., et al., *Validity and Reliability of Field-Based Measures for Assessing Movement Skill Competency in Lifelong Physical Activities: A Systematic Review.* Sports medicine (Auckland, N.Z.), 2015. **45**(10): p. 1443-54.
299. Janaudis-Ferreira, T., et al., *How should we measure arm exercise capacity in COPD? A systematic review.* European Respiratory Journal, 2011. **38**(SUPPL. 55).
300. Jaspers, M.E.H., et al., *A systematic review on the quality of measurement techniques for the assessment of burn wound depth or healing potential.* Burns, 2019. **45**(2): p. 261-281.
301. Johnston, K.N., A.J. Potter, and A. Phillips, *Measurement Properties of Short Lower Extremity Functional Exercise Tests in People With Chronic Obstructive Pulmonary Disease: Systematic Review.* Physical therapy, 2017. **97**(9): p. 926-943.
302. Jonsson, A. and E. Rasmussen-Barr, *Intra- and inter-rater reliability of movement and palpation tests in patients with neck pain: A systematic review.* Physiotherapy theory and practice, 2018. **34**(3): p. 165-180.
303. Jorgensen, L.B., et al., *Methods to assess area and volume of wounds - a systematic review.* International wound journal, 2016. **13**(4): p. 540-53.
304. Kasehagen, B., et al., *Assessing the Reliability of Ultrasound Imaging to Examine Peripheral Nerve Excursion: A Systematic Literature Review.* Ultrasound in medicine & biology, 2018. **44**(1): p. 1-13.
305. Klingels, K., et al., *A systematic review of arm activity measures for children with hemiplegic cerebral palsy.* Clinical rehabilitation, 2010. **24**(10): p. 887-900.
306. Konieczka, C., et al., *What is the reliability of clinical measurement tests for humeral head position? A systematic review.* Journal of hand therapy : official journal of the American Society of Hand Therapists, 2017. **30**(4): p. 420-431.
307. Kristensen, O.H., E. Stenager, and U. Dalgas, *Muscle Strength and Poststroke Hemiplegia: A Systematic Review of Muscle Strength Assessment and Muscle Strength Impairment.* Archives of physical medicine and rehabilitation, 2017. **98**(2): p. 368-380.
308. Kroman, S.L., et al., *Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: a systematic review.* Osteoarthritis and cartilage, 2014. **22**(1): p. 26-39.
309. Kwah, L.K., et al., *Reliability and validity of ultrasound measurements of muscle fascicle length and pennation in humans: a systematic review.* Journal of applied physiology (Bethesda, Md. : 1985), 2013. **114**(6): p. 761-9.
310. Lagarde, M.L.J., D.M.A. Kamalski, and L. van den Engel-Hoek, *The reliability and validity of cervical auscultation in the diagnosis of dysphagia: a systematic review.* Clinical rehabilitation, 2016. **30**(2): p. 199-207.
311. Lamers, I., et al., *Upper limb assessment in multiple sclerosis: a systematic review of outcome measures and their psychometric properties.* Archives of physical medicine and rehabilitation, 2014. **95**(6): p. 1184-200.
312. Lange, T., et al., *The reliability of physical examination tests for the clinical assessment of scapular dyskinesis in subjects with shoulder complaints: A systematic review.* Physical therapy in sport : official journal of the Association of Chartered Physiotherapists in Sports Medicine, 2017. **26**: p. 64-89.
313. Larsen, C.M., et al., *Measurement properties of existing clinical assessment methods evaluating scapular positioning and function. A systematic review.* Physiotherapy theory and practice, 2014. **30**(7): p. 453-82.
314. Le Flao, E., et al., *Assessing Head/Neck Dynamic Response to Head Perturbation: A Systematic Review.* Sports medicine (Auckland, N.Z.), 2018. **48**(11): p. 2641-2658.
315. Lemeunier, N., et al., *Reliability and validity of clinical tests to assess the anatomical integrity of the cervical spine in adults with neck pain and its associated disorders: Part 1-A systematic review from the Cervical Assessment and Diagnosis Research Evaluation (CADRE)*

Collaboration. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 2017. **26**(9): p. 2225-2241.
316. Lemeunier, N., et al., *Reliability and validity of clinical tests to assess posture, pain location, and cervical spine mobility in adults with neck pain and its associated disorders: Part 4. A systematic review from the cervical assessment and diagnosis research evaluation (CADRE) collaboration.* Musculoskeletal Science and Practice, 2018. **38**: p. 128-147.
317. Lennon, N., et al., *The clinimetric properties of aerobic and anaerobic fitness measures in adults with cerebral palsy: A systematic review of the literature.* Research in developmental disabilities, 2015. **45-46**: p. 316-28.
318. Liang, Q. and M. Sarkar, *Intra- and inter-individual variability in urinary nicotine excretion and plasma cotinine in adult cigarette smokers.* Regulatory toxicology and pharmacology : RTP, 2012. **64**(3): p. 388-93.
319. Andrade Lima, C., et al., *Six-minute walk test as a determinant of the functional capacity of children and adolescents with cystic fibrosis: A systematic review.* Respir Med, 2018. **137**: p. 83-88.
320. Lisboa, C.d.O., et al., *Reliability and reproducibility of three-dimensional cephalometric landmarks using CBCT: a systematic review.* Journal of applied oral science : revista FOB, 2015. **23**(2): p. 112-9.
321. Maaswinkel, E., et al., *Methods for assessment of trunk stabilization, a systematic review.* Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology, 2016. **26**: p. 18-35.
322. MacKay, J.W., et al., *Systematic review and meta-analysis of the reliability and discriminative validity of cartilage compositional MRI in knee osteoarthritis.* Osteoarthritis and Cartilage, 2018. **26**(9): p. 1140-1152.
323. Mahaffey, R., et al., *Clinical outcome measures for monitoring physical function in pediatric obesity: An integrative review.* Obesity (Silver Spring, Md.), 2016. **24**(5): p. 993-1017.
324. Maricar, N., et al., *Clinical assessment of effusion in knee osteoarthritis-A systematic review.* Seminars in arthritis and rheumatism, 2016. **45**(5): p. 556-63.
325. Marshall, N., E. Ward, and C.M. Williams, *The identification and appraisal of assessment tools used to evaluate metatarsus adductus: a systematic review of their measurement properties.* Journal of foot and ankle research, 2018. **11**: p. 25.
326. May, S., et al., *Reliability of physical examination tests used in the assessment of patients with shoulder problems: a systematic review.* Physiotherapy, 2010. **96**(3): p. 179-90.
327. Mc Auliffe, S., et al., *A systematic review of the reliability of diagnostic ultrasound imaging in measuring tendon size: Is the error clinically acceptable?* Physical therapy in sport : official journal of the Association of Chartered Physiotherapists in Sports Medicine, 2017. **26**: p. 52-63.
328. McCreesh, K.M., J.M. Crotty, and J.S. Lewis, *Acromiohumeral distance measurement in rotator cuff tendinopathy: is there a reliable, clinically applicable method? A systematic review.* British journal of sports medicine, 2015. **49**(5): p. 298-305.
329. McVerry, F., D.S. Liebeskind, and K.W. Muir, *Systematic review of methods for assessing leptomeningeal collateral flow.* AJNR. American journal of neuroradiology, 2012. **33**(3): p. 576-82.
330. Michiels, S., et al., *The assessment of cervical sensory motor control: a systematic review focusing on measuring methods and their clinimetric characteristics.* Gait & posture, 2013. **38**(1): p. 1-7.
331. Mieritz, R.M., et al., *Reliability and measurement error of 3-dimensional regional lumbar motion measures: a systematic review.* Journal of manipulative and physiological therapeutics, 2012. **35**(8): p. 645-56.

332. Mijnarends, D.M., et al., *Validity and reliability of tools to measure muscle mass, strength, and physical performance in community-dwelling older people: a systematic review.* Journal of the American Medical Directors Association, 2013. **14**(3): p. 170-8.
333. Milani, P., et al., *Mobile smartphone applications for body position measurement in rehabilitation: a review of goniometric tools.* PM & R : the journal of injury, function, and rehabilitation, 2014. **6**(11): p. 1038-43.
334. Milne, S.C., et al., *Psychometric properties of outcome measures evaluating decline in gait in cerebellar ataxia: A systematic review.* Gait & posture, 2018. **61**: p. 149-162.
335. Mohan, V., et al., *Reliability of diaphragmatic mobility assessment: A systematic review.* Polish Annals of Medicine, 2018. **25**(2): p. 266-271.
336. Mohseni Bandpei, M.A., et al., *Reliability of surface electromyography in the assessment of paraspinal muscle fatigue: an updated systematic review.* Journal of manipulative and physiological therapeutics, 2014. **37**(7): p. 510-21.
337. Moloney, N.A., T.M. Hall, and C.M. Doody, *Reliability of thermal quantitative sensory testing: a systematic review.* Journal of rehabilitation research and development, 2012. **49**(2): p. 191-207.
338. Moore, M. and K. Barker, *The validity and reliability of the four square step test in different adult populations: a systematic review.* Systematic reviews, 2017. **6**(1): p. 187.
339. Mulder-Brouwer, A.N., E.A.A. Rameckers, and C.H. Bastiaenen, *Lower Extremity Handheld Dynamometry Strength Measurement in Children With Cerebral Palsy.* Pediatric physical therapy : the official publication of the Section on Pediatrics of the American Physical Therapy Association, 2016. **28**(2): p. 136-53.
340. Muntaner-Mas, A., et al., *A Systematic Review of Fitness Apps and Their Potential Clinical and Sports Utility for Objective and Remote Assessment of Cardiorespiratory Fitness.* Sports medicine (Auckland, N.Z.), 2019. **49**(4): p. 587-600.
341. Nae, J., et al., *Measurement properties of visual rating of postural orientation errors of the lower extremity - A systematic review and meta-analysis.* Physical therapy in sport : official journal of the Association of Chartered Physiotherapists in Sports Medicine, 2017. **27**: p. 52-64.
342. Coelho Neto, M.A., et al., *True Reproducibility of UltraSound Techniques (TRUST): systematic review of reliability studies in obstetrics and gynecology.* Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology, 2015. **46**(1): p. 14-20.
343. Nijholt, W., et al., *The reliability and validity of ultrasound to quantify muscles in older adults: a systematic review.* Journal of cachexia, sarcopenia and muscle, 2017. **8**(5): p. 702-712.
344. Oberwahrenbrock, T., et al., *Reliability of Intra-Retinal Layer Thickness Estimates.* PloS one, 2015. **10**(9): p. e0137316.
345. O'Meara, S.M., et al., *Systematic review of wound measurement instruments.* Wound Repair and Regeneration, 2012. **20**(3): p. 263-276.
346. Ornetti, P., et al., *Gait analysis as a quantifiable outcome measure in hip or knee osteoarthritis: a systematic review.* Joint, bone, spine : revue du rhumatisme, 2010. **77**(5): p. 421-5.
347. Ortega, F.B., et al., *Systematic review and proposal of a field-based physical fitness-test battery in preschool children: the PREFIT battery.* Sports medicine (Auckland, N.Z.), 2015. **45**(4): p. 533-55.
348. Ozay, H., A. Cakir, and M.C. Ecevit, *Retronasal Olfaction Test Methods: A Systematic Review.* Balkan medical journal, 2019. **36**(1): p. 49-59.
349. Paech, D.C., et al., *A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer.* Journal of

thoracic oncology : official publication of the International Association for the Study of Lung Cancer, 2011. **6**(1): p. 55-63.
350. Parreira, V.F., et al., *Measurement properties of the incremental shuttle walk test. a systematic review.* Chest, 2014. **145**(6): p. 1357-1369.
351. Parry, S.M., et al., *Assessment of impairment and activity limitations in the critically ill: a systematic review of measurement instruments and their clinimetric properties.* Intensive care medicine, 2015. **41**(5): p. 744-62.
352. Paul, D.J., T.J. Gabbett, and G.P. Nassis, *Agility in Team Sports: Testing, Training and Factors Affecting Performance.* Sports medicine (Auckland, N.Z.), 2016. **46**(3): p. 421-42.
353. Peeling, R.W., et al., *CD4 enumeration technologies: a systematic review of test performance for determining eligibility for antiretroviral therapy.* PloS one, 2015. **10**(3): p. e0115019.
354. Perdomo, M., et al., *Assessment measures of secondary lymphedema in Breast Cancer survivors.* Rehabilitation Oncology, 2014. **32**(1): p. 22-35.
355. Petitclerc, E., et al., *Lower limb muscle impairment in myotonic dystrophy type 1: the need for better guidelines.* Muscle & nerve, 2015. **51**(4): p. 473-8.
356. Phillips, S.S., et al., *A Systematic Review Assessing the Current State of Automated Pupillometry in the NeuroICU.* Neurocritical Care, 2018.
357. Pin, T.W., *Psychometric properties of 2-minute walk test: a systematic review.* Archives of physical medicine and rehabilitation, 2014. **95**(9): p. 1759-75.
358. Pollock, C., J. Eng, and S. Garland, *Clinical measurement of walking balance in people post stroke: a systematic review.* Clinical rehabilitation, 2011. **25**(8): p. 693-708.
359. Ponce-Garcia, C., et al., *Reliability of three-dimensional anterior cranial base superimposition methods for assessment of overall hard tissue changes: A systematic review.* The Angle orthodontist, 2018. **88**(2): p. 233-245.
360. Pons, C., et al., *Validity and reliability of radiological methods to assess proximal hip geometry in children with cerebral palsy: a systematic review.* Developmental medicine and child neurology, 2013. **55**(12): p. 1089-102.
361. Pons, C., et al., *Quantifying skeletal muscle volume and shape in humans using MRI: A systematic review of validity and reliability.* PLoS ONE, 2018. **13**(11): p. e0207847.
362. Proud, E.L., et al., *Evaluation of measures of upper limb functioning and disability in people with Parkinson disease: a systematic review.* Archives of physical medicine and rehabilitation, 2015. **96**(3): p. 540-551.e1.
363. Prowse, A., et al., *Reliability and validity of inexpensive and easily administered anthropometric clinical evaluation methods of postural asymmetry measurement in adolescent idiopathic scoliosis: a systematic review.* European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 2016. **25**(2): p. 450-66.
364. Ratter, J., L. Radlinger, and C. Lucas, *Several submaximal exercise tests are reliable, valid and acceptable in people with chronic pain, fibromyalgia or chronic fatigue: a systematic review.* Journal of physiotherapy, 2014. **60**(3): p. 144-50.
365. Reis Durao, A.P., et al., *Masseter muscle measurement performed by ultrasound: a systematic review.* Dento maxillo facial radiology, 2017. **46**(6): p. 20170052.
366. Ringshausen, F.C., A. Schablon, and A. Nienhaus, *Interferon-gamma release assays for the tuberculosis serial testing of health care workers: A systematic review.* Journal of Occupational Medicine and Toxicology, 2012. **7**(1): p. 6.
367. Roberts, H.C., et al., *A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach.* Age and ageing, 2011. **40**(4): p. 423-9.
368. Robertson, S.J., A.F. Burnett, and J. Cochrane, *Tests examining skill outcomes in sport: a systematic review of measurement properties and feasibility.* Sports medicine (Auckland, N.Z.), 2014. **44**(4): p. 501-18.

369. Roeing, K.L., K.L. Hsieh, and J.J. Sosnoff, *A systematic review of balance and fall risk assessments with mobile phone technology.* Archives of gerontology and geriatrics, 2017. **73**: p. 222-226.
370. Rondoni, A., et al., *Intrarater and Inter-rater Reliability of Active Cervical Range of Motion in Patients With Nonspecific Neck Pain Measured With Technological and Common Use Devices: A Systematic Review With Meta-regression.* Journal of manipulative and physiological therapeutics, 2017. **40**(8): p. 597-608.
371. Rossini, G., et al., *Diagnostic accuracy and measurement sensitivity of digital models for orthodontic purposes: A systematic review.* American journal of orthodontics and dentofacial orthopedics : official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics, 2016. **149**(2): p. 161-70.
372. Rubio-Ochoa, J., et al., *Physical examination tests for screening and diagnosis of cervicogenic headache: A systematic review.* Manual therapy, 2016. **21**: p. 35-40.
373. Ruhe, A., R. Fejer, and B. Walker, *The test-retest reliability of centre of pressure measures in bipedal static task conditions--a systematic review of the literature.* Gait & posture, 2010. **32**(4): p. 436-45.
374. Rydwik, E., et al., *Investigation into the reliability and validity of the measurement of elderly people's clinical walking speed: a systematic review.* Physiotherapy theory and practice, 2012. **28**(3): p. 238-56.
375. Saccomanno, M.F., et al., *Magnetic resonance imaging criteria for the assessment of the rotator cuff after repair: a systematic review.* Knee surgery, sports traumatology, arthroscopy : official journal of the ESSKA, 2015. **23**(2): p. 423-42.
376. Saether, R., et al., *Clinical tools to assess balance in children and adults with cerebral palsy: a systematic review.* Developmental medicine and child neurology, 2013. **55**(11): p. 988-99.
377. Saltzherr, M.S., et al., *Metric properties of advanced imaging methods in osteoarthritis of the hand: a systematic review.* Annals of the rheumatic diseases, 2014. **73**(2): p. 365-75.
378. Sam, A., et al., *Reliability of different three-dimensional cephalometric landmarks in cone-beam computed tomography : A systematic review.* The Angle orthodontist, 2019. **89**(2): p. 317-332.
379. Scalco, J.C., et al., *Psychometric properties of functional capacity tests in children and adolescents: Systematic review.* Revista Paulista de Pediatria, 2018. **36**(4): p. 500-510.
380. Scheetz, J., et al., *Validity and reliability of eye healthcare professionals in the assessment of glaucoma - a systematic review.* International journal of clinical practice, 2015. **69**(6): p. 689-702.
381. Schrama, P.P.M., et al., *Intraexaminer reliability of hand-held dynamometry in the upper extremity: a systematic review.* Archives of physical medicine and rehabilitation, 2014. **95**(12): p. 2444-69.
382. Seager, A., H. French, and D. Meldrum, *Measurement properties of instruments for assessment of cervical spine function in infants with torticollis: a systematic review.* European Journal of Pediatrics, 2019.
383. Sepriano, A., et al., *DXA in the assessment of subchondral bone mineral density in knee osteoarthritis--A semi-standardized protocol after systematic review.* Seminars in arthritis and rheumatism, 2015. **45**(3): p. 275-83.
384. Shiel, F., et al., *Dual energy X-ray absorptiometry positioning protocols in assessing body composition: A systematic review of the literature.* Journal of science and medicine in sport, 2018. **21**(10): p. 1038-1044.
385. Silva, P.F.S., et al., *Measurement properties and feasibility of clinical tests to assess sit-to-stand/stand-to-sit tasks in subjects with neurological disease: a systematic review.* Brazilian journal of physical therapy, 2014. **18**(2): p. 99-110.

386. Simperingham, K.D., J.B. Cronin, and A. Ross, *Advances in Sprint Acceleration Profiling for Field-Based Team-Sport Athletes: Utility, Reliability, Validity and Limitations.* Sports medicine (Auckland, N.Z.), 2016. **46**(11): p. 1619-1645.
387. Singh, S.J., et al., *An official systematic review of the European Respiratory Society/American Thoracic Society: Measurement properties of field walking tests in chronic respiratory disease.* European Respiratory Journal, 2014. **44**(6): p. 1447-1478.
388. Sman, A.D., C.E. Hiller, and K.M. Refshauge, *Diagnostic accuracy of clinical tests for diagnosis of ankle syndesmosis injury: a systematic review.* British journal of sports medicine, 2013. **47**(10): p. 620-8.
389. Smith, T.O., L. Davies, and C.B. Hing, *A systematic review to determine the reliability of knee joint position sense assessment measures.* The Knee, 2013. **20**(3): p. 162-9.
390. Smith, T.O., et al., *The reliability and validity of radiological assessment for patellar instability. A systematic review and meta-analysis.* Skeletal radiology, 2011. **40**(4): p. 399-414.
391. Sollis, K.A., et al., *Systematic review of the performance of HIV viral load technologies on plasma samples.* PloS one, 2014. **9**(2): p. e85869.
392. Southerst, D., et al., *The reliability of body pain diagrams in the quantitative measurement of pain distribution and location in patients with musculoskeletal pain: a systematic review.* Journal of manipulative and physiological therapeutics, 2013. **36**(7): p. 450-9.
393. Stephensen, D., W.I. Drechsler, and O.M. Scott, *Outcome measures monitoring physical function in children with haemophilia: a systematic review.* Haemophilia : the official journal of the World Federation of Hemophilia, 2014. **20**(3): p. 306-21.
394. Stienen, M.N., et al., *Objective measures of functional impairment for degenerative diseases of the lumbar spine: a systematic review of the literature.* Spine Journal, 2019.
395. Stovall, B.A. and S. Kumar, *Reliability of bony anatomic landmark asymmetry assessment in the lumbopelvic region: application to osteopathic medical education.* The Journal of the American Osteopathic Association, 2010. **110**(11): p. 667-74.
396. Symonds, T., P. Campbell, and J.A. Randall, *A review of muscle- and performance-based assessment instruments in DM1.* Muscle and Nerve, 2017. **56**(1): p. 78-85.
397. Talma, H., et al., *Bioelectrical impedance analysis to estimate body composition in children and adolescents: a systematic review and evidence appraisal of validity, responsiveness, reliability and measurement error.* Obesity reviews : an official journal of the International Association for the Study of Obesity, 2013. **14**(11): p. 895-905.
398. Tarara, D.T., et al., *Clinician-friendly physical performance tests in athletes part 3: a systematic review of measurement properties and correlations to injury for tests in the upper extremity.* British journal of sports medicine, 2016. **50**(9): p. 545-51.
399. Terwee, C.B., et al., *Instruments to assess physical activity in patients with osteoarthritis of the hip or knee: a systematic review of measurement properties.* Osteoarthritis and cartilage, 2011. **19**(6): p. 620-33.
400. Timmer, M.A., et al., *Measuring activities and participation in persons with haemophilia: A systematic review of commonly used instruments.* Haemophilia : the official journal of the World Federation of Hemophilia, 2018. **24**(2): p. e33-e49.
401. Tong, L. and L.S. Teng, *Review of Literature on Measurements of Non-invasive Break Up Times, Lipid Morphology and Tear Meniscal Height Using Commercially Available Hand-held Instruments.* Current Eye Research, 2018. **43**(5): p. 567-575.
402. Toohey, L.A., et al., *Is a sphygmomanometer a valid and reliable tool to measure the isometric strength of hip muscles? A systematic review.* Physiotherapy theory and practice, 2015. **31**(2): p. 114-9.
403. Traverso, A., et al., *Repeatability and Reproducibility of Radiomic Features: A Systematic Review.* International Journal of Radiation Oncology Biology Physics, 2018. **102**(4): p. 1143-1158.

404. Valet, M., et al., *Quality of the tools used to assess aerobic capacity in people with multiple sclerosis.* European journal of physical and rehabilitation medicine, 2017. **53**(5): p. 759-774.
405. van Bloemendaal, M., A.T.M. van de Water, and I.G.L. van de Port, *Walking tests for stroke survivors: a systematic review of their measurement properties.* Disability and rehabilitation, 2012. **34**(26): p. 2207-21.
406. van de Pol, R.J., E. van Trijffel, and C. Lucas, *Inter-rater reliability for measurement of passive physiological range of motion of upper extremity joints is better if instruments are used: a systematic review.* Journal of physiotherapy, 2010. **56**(1): p. 7-17.
407. van de Water, A.T.M. and D.R. Benjamin, *Measurement methods to assess diastasis of the rectus abdominis muscle (DRAM): A systematic review of their measurement properties and meta-analytic reliability generalisation.* Manual therapy, 2016. **21**: p. 41-53.
408. van Kooij, Y.E., et al., *The reliability and measurement error of protractor-based goniometry of the fingers: A systematic review.* Journal of hand therapy : official journal of the American Society of Hand Therapists, 2017. **30**(4): p. 457-467.
409. van Trijffel, E., et al., *Inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low: a systematic review.* Journal of physiotherapy, 2010. **56**(4): p. 223-35.
410. Wang, K.C., et al., *Evidence-based outcomes on diagnostic accuracy of quantitative ultrasound for assessment of pediatric osteoporosis - a systematic review.* Pediatric radiology, 2014. **44**(12): p. 1573-87.
411. Wen, D., et al., *Measurement properties and feasibility of the Loughborough soccer passing test: A systematic review.* Journal of sports sciences, 2018. **36**(15): p. 1682-1694.
412. Wetterslev, M., et al., *Systematic review of cardiac output measurements by echocardiography vs. thermodilution: the techniques are not interchangeable.* Intensive care medicine, 2016. **42**(8): p. 1223-33.
413. Williams, M.A., et al., *A Systematic Review of Reliability and Validity Studies of Methods for Measuring Active andPassive Cervical Range of Motion.* Journal of Manipulative and Physiological Therapeutics, 2010. **33**(2): p. 138-155.
414. Winser, S.J., et al., *Systematic review of the psychometric properties of balance measures for cerebellar ataxia.* Clinical rehabilitation, 2015. **29**(1): p. 69-79.
415. Wouters, M., H.M. Evenhuis, and T.I.M. Hilgenkamp, *Systematic review of field-based physical fitness tests for children and adolescents with intellectual disabilities.* Research in developmental disabilities, 2017. **61**: p. 77-94.
416. Yang, L., et al., *Psychometric properties of dual-task balance and walking assessments for individuals with neurological conditions: A systematic review.* Gait & posture, 2017. **52**: p. 110-123.
417. Yang, L., et al., *Psychometric properties of dual-task balance assessments for older adults: a systematic review.* Maturitas, 2015. **80**(4): p. 359-69.
418. Zanudin, A., et al., *Psychometric properties of measures of gait quality and walking performance in young people with Cerebral Palsy: A systematic review.* Gait & posture, 2017. **58**: p. 30-40.
419. Zayat, A.S., et al., *The role of ultrasound in assessing musculoskeletal symptoms of systemic lupus erythematosus: a systematic literature review.* Rheumatology (Oxford, England), 2016. **55**(3): p. 485-94.
420. Zimmerman, J.N., J. Lee, and B.T. Pliska, *Reliability of upper pharyngeal airway assessment using dental CBCT: a systematic review.* European journal of orthodontics, 2017. **39**(5): p. 489-496.
421. Terwee, C.B., et al., *Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist.* Qual Life Res, 2012. **21**(4): p. 651-7.

422. Whiting, P., et al., *The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews.* BMC Medical Research Methodology, 2003. **3**(1): p. 25.
423. Lucas, N.P., et al., *The development of a quality appraisal tool for studies of diagnostic reliability (QAREL).* J Clin Epidemiol, 2010. **63**(8): p. 854-61.
424. Kmet, L. and R. Lee, *Standard Quality Assessment Criteria for Evaluating Primary Research Papers from a Variety of FieldsAHFMRHTA Initiative20040213.* HTA Initiative, 2004. **2**.
425. Hollis-Sawyer, L., *Evidence-Based Rehabilitation: A Guide to Practice, by Mary Law and Joy MacDermid.* Activities, Adaptation & Aging, 2017. **41**(2): p. 193-194.
426. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
427. Essendrop, M., et al., *Measures of low back function: a review of reproducibility studies.* Clin Biomech (Bristol, Avon), 2002. **17**(4): p. 235-49.
428. Jadad, A.R., et al., *Assessing the quality of reports of randomized clinical trials: is blinding necessary?* Control Clin Trials, 1996. **17**(1): p. 1-12.
429. Brink, Y. and Q.A. Louw, *Clinical instruments: reliability and validity critical appraisal.* J Eval Clin Pract, 2012. **18**(6): p. 1126-32.
430. Chipchase, L., et al., *A checklist for assessing the methodological quality of studies using transcranial magnetic stimulation to study the motor system: An international consensus study.* Clinical Neurophysiology, 2012. **123**(9): p. 1698-1704.
431. Harrington, S., L. Gilchrist, and A. Sander, *Breast Cancer EDGE Task Force Outcomes: Clinical Measures of Pain.* Rehabil Oncol, 2014. **32**(1): p. 13-21.
432. Bialocerkowski, A., N. Klupp, and P. Bragge, *How to read and critically appraise a reliability article.* International Journal of Therapy and Rehabilitation, 2010. **17**: p. 114-120.
433. Robertson, S.J., A.F. Burnett, and J. Cochrane, *Tests Examining Skill Outcomes in Sport: A Systematic Review of Measurement Properties and Feasibility.* Sports Medicine, 2014. **44**(4): p. 501-518.
434. Downs, S.H. and N. Black, *The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions.* Journal of Epidemiology and Community Health, 1998. **52**(6): p. 377-384.
435. Viswanathan, M., et al., *AHRQ Methods for Effective Health Care. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions*, in *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. 2008, Agency for Healthcare Research and Quality (US): Rockville (MD).
436. Pretorius, A. and J.L. Keating, *Validity of real time ultrasound for measuring skeletal muscle size.* Physical Therapy Reviews, 2008. **13**(6): p. 415-426.
437. Whiting, P.F., et al., *QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.* Ann Intern Med, 2011. **155**(8): p. 529-36.
438. Bossuyt, P.M., et al., *STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies.* Bmj, 2015. **351**: p. h5527.
439. Hebert, J., et al., *A Systematic Review of the Reliability of Rehabilitative Ultrasound Imaging for the Quantitative Assessment of the Abdominal and Lumbar Trunk Muscles.* Spine, 2009. **34**: p. E848-56.
440. Vermeulen, H., S. Berben, and M. Heinen, *Evidence based richtlijnen, werken volgens de laatste stand van wetenschap.* JGZ Tijdschrift voor jeugdgezondheidszorg, 2017. **49**(4): p. 67-68.
441. von Elm, E., et al., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies.* The Lancet, 2007. **370**(9596): p. 1453-1457.
442. van Vlijmen, O.J., et al., *Evidence supporting the use of cone-beam computed tomography in orthodontics.* J Am Dent Assoc, 2012. **143**(3): p. 241-52.

443. Higgins, J.P., et al., *The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.* Bmj, 2011. **343**: p. d5928.
444. Merlin, T., A. Weston, and R. Tooher, *Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'.* BMC Med Res Methodol, 2009. **9**: p. 34.
445. Brughelli, M., et al., *Understanding Change of Direction Ability in Sport.* Sports Medicine, 2008. **38**(12): p. 1045-1063.
446. van Bloemendaal, M., A.T.M. van de Water, and I.G.L. van de Port, *Walking tests for stroke survivors: a systematic review of their measurement properties.* Disability and Rehabilitation, 2012. **34**(26): p. 2207-2221.
447. Bossuyt, P.M., et al., *Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative.* Bmj, 2003. **326**(7379): p. 41-4.
448. Alsufyani, N.A., C. Flores-Mir, and P.W. Major, *Three-dimensional segmentation of the upper airway using cone beam CT: a systematic review.* Dentomaxillofac Radiol, 2012. **41**(4): p. 276-84.
449. Abizanda, P., et al., *Validity and usefulness of hand-held dynamometry for measuring muscle strength in community-dwelling older persons.* Archives of Gerontology and Geriatrics, 2012. **54**(1): p. 21-27.
450. Aguiar, L.T., et al., *Dynamometry for the measurement of grip, pinch, and trunk muscles strength in subjects with subacute stroke: reliability and different number of trials.* Brazilian Journal of Physical Therapy / Revista Brasileira de Fisioterapia, 2016. **20**(5): p. 395-404.
451. Allen, D., et al., *Reliability and validity of an electronic dynamometer for measuring grip strength...including commentary by Stapanian MA and Fitinghoff H.* International Journal of Therapy & Rehabilitation, 2011. **18**(5): p. 258-265.
452. Anumula, S.K., C. Beku, and Y.S.N. Murthy, *Measurement of Reliability in Grip Strength.* Indian Journal of Physiotherapy & Occupational Therapy, 2014. **8**(2): p. 115-119.
453. Baldwin, C.E., J.D. Paratz, and A.D. Bersten, *Muscle strength assessment in critically ill patients with handheld dynamometry: An investigation of reliability, minimal detectable change, and time to peak force generation.* Journal of Critical Care, 2013. **28**(1): p. 77-86.
454. Bertrand, A.M., et al., *Reliability of maximal grip strength measurements and grip strength recovery following a stroke.* Journal of Hand Therapy, 2015. **28**(4): p. 356-363.
455. Blankevoort, C.G., M.J.G. van Heuvelen, and E.J.A. Scherder, *Reliability of Six Physical Performance Tests in Older People With Dementia.* Physical Therapy, 2013. **93**(1): p. 69-78.
456. Bodilsen, A.C., et al., *Feasibility and inter-rater reliability of physical performance measures in acutely admitted older medical patients.* PLoS ONE, 2015. **10 (2) (no pagination)**(e0118248).
457. Bohannon, R.W., *Parallel comparison of grip strength measures obtained with A microfet 4 and A jamar dynamometer.* Perceptual and Motor Skills, 2005. **100**(3 I): p. 795-798.
458. Bohannon, R.W., *Test-retest reliability of the MicroFET 4 hand-grip dynamometer.* Physiotherapy Theory & Practice, 2006. **22**(4): p. 219-221.
459. Bohannon, R.W., *Test-retest reliability of the five-repetition sit-to-stand test: a systematic review of the literature involving adults.* Journal of Strength & Conditioning Research (Lippincott Williams & Wilkins), 2011. **25**(11): p. 3205-3207.
460. Boissy, P., et al., *Maximal grip force in chronic stroke subjects and its relationship to global upper extremity function.* Clinical Rehabilitation, 1999. **13**(4): p. 354-362.
461. Brogårdh, C., et al., *Intra-rater Reliability of Arm and Hand Muscle Strength Measurements in Persons With Late Effects of Polio.* PM & R: Journal of Injury, Function & Rehabilitation, 2015. **7**(10): p. 1035-1041.
462. Brown, A., et al., *Validity and reliability of the Dexter Hand Evaluation and Therapy System in hand-injured patients.* Journal of Hand Therapy, 2000. **13**(1): p. 37-45.

463. Buehring, B., et al., *Reproducibility of jumping mechanography and traditional measures of physical and muscle function in older adults.* Osteoporosis International, 2014. **26**(2): p. 819-825.
464. Burnstein, B.D., R.J. Steele, and I. Shrier, *Reliability of Fitness Tests Using Methods and Time Periods Common in Sport and Occupational Management.* Journal of Athletic Training (National Athletic Trainers' Association), 2011. **46**(5): p. 505-513.
465. Carbonell-Baeza, A., et al., *Reliability and Feasibility of Physical Fitness Tests in Female Fibromyalgia Patients.* International Journal of Sports Medicine, 2015. **36**(2): p. 157-162.
466. Chen, H., et al., *Test-retest reproducibility and smallest real difference of 5 hand function tests in patients with stroke.* Neurorehabilitation & Neural Repair, 2009. **23**(5): p. 435-440.
467. Ekstrand, E., J. Lexell, and C. Brogardh, *Isometric and isokinetic muscle strength in the upper extremity can be reliably measured in persons with chronic stroke.* Journal of rehabilitation medicine, 2015. **47**(8): p. 706-713.
468. Essendrop, M., B. Schibye, and K. Hansen, *Reliability of isometric muscle strength tests for the trunk, hands and shoulders.* International Journal of Industrial Ergonomics, 2001. **28**(6): p. 379-387.
469. Fox, B., et al., *Relative and absolute reliability of functional performance measures for adults with dementia living in residential aged care.* International Psychogeriatrics, 2014. **26**(10): p. 1659-1667.
470. Gatt, I., et al., *The Takei Handheld Dynamometer: An Effective Clinical Outcome Measure Tool for Hand and Wrist Function in Boxing.* Hand (New York, N.Y.), 2018. **13**(3): p. 319-324.
471. Gerhardsson, L., L. Gillstrom, and M. Hagberg, *Test-retest reliability of neurophysiological tests of hand-arm vibration syndrome in vibration exposed workers and unexposed referents.* Journal of Occupational Medicine and Toxicology, 2014. **9 (1) (no pagination)**(38).
472. Gerodimos, V., *Reliability of handgrip strength test in basketball players.* Journal of Human Kinetics, 2012. **31**: p. 25-36.
473. Gittings, P., et al., *Grip and Muscle Strength Dynamometry Are Reliable and Valid in Patients With Unhealed Minor Burn Wounds.* Journal of Burn Care & Research, 2016. **37**(6): p. 388-396.
474. Gittings, P.M., et al., *Grip and Muscle Strength Dynamometry in Acute Burn Injury: Evaluation of an Updated Assessment Protocol.* Journal of Burn Care & Research, 2018. **39**(6): p. 937-947.
475. Haidar, S.G., et al., *Average versus maximum grip strength: Which is more consistent?* Journal of Hand Surgery, 2004. **29 B**(1): p. 82-84.
476. Haward, B.M. and M.J. Griffin, *Repeatability of grip strength and dexterity tests and the effects of age and gender.* International Archives of Occupational and Environmental Health, 2002. **75**(1-2): p. 111-119.
477. Hilgenkamp, T.I.M., R. van Wijck, and H.M. Evenhuis, *Feasibility and reliability of physical fitness tests in older adults with intellectual disability: A pilot study.* Journal of Intellectual & Developmental Disability, 2012. **37**(2): p. 158-162.
478. Irwin, C.B. and M.E. Sesto, *Reliability and validity of the multiaxis profile dynamometer with younger and older participants.* Journal of Hand Therapy, 2010. **23**(3): p. 281-289.
479. Jenkins, N.D.M. and J.T. Cramer, *Reliability and Minimum Detectable Change for Common Clinical Physical Function Tests in Sarcopenic Men and Women.* Journal of the American Geriatrics Society, 2017. **65**(4): p. 839-846.
480. Kennedy, D., C. Jerosch-Herold, and M. Hickson, *The reliability of one vs. three trials of pain-free grip strength in subjects with rheumatoid arthritis.* Journal of Hand Therapy, 2010. **23**(4): p. 384-[436].
481. Khamwong, P., et al., *Reliability of muscle function and sensory perception measurements of the wrist extensors.* Physiotherapy theory and practice, 2010. **26**(6): p. 408-415.

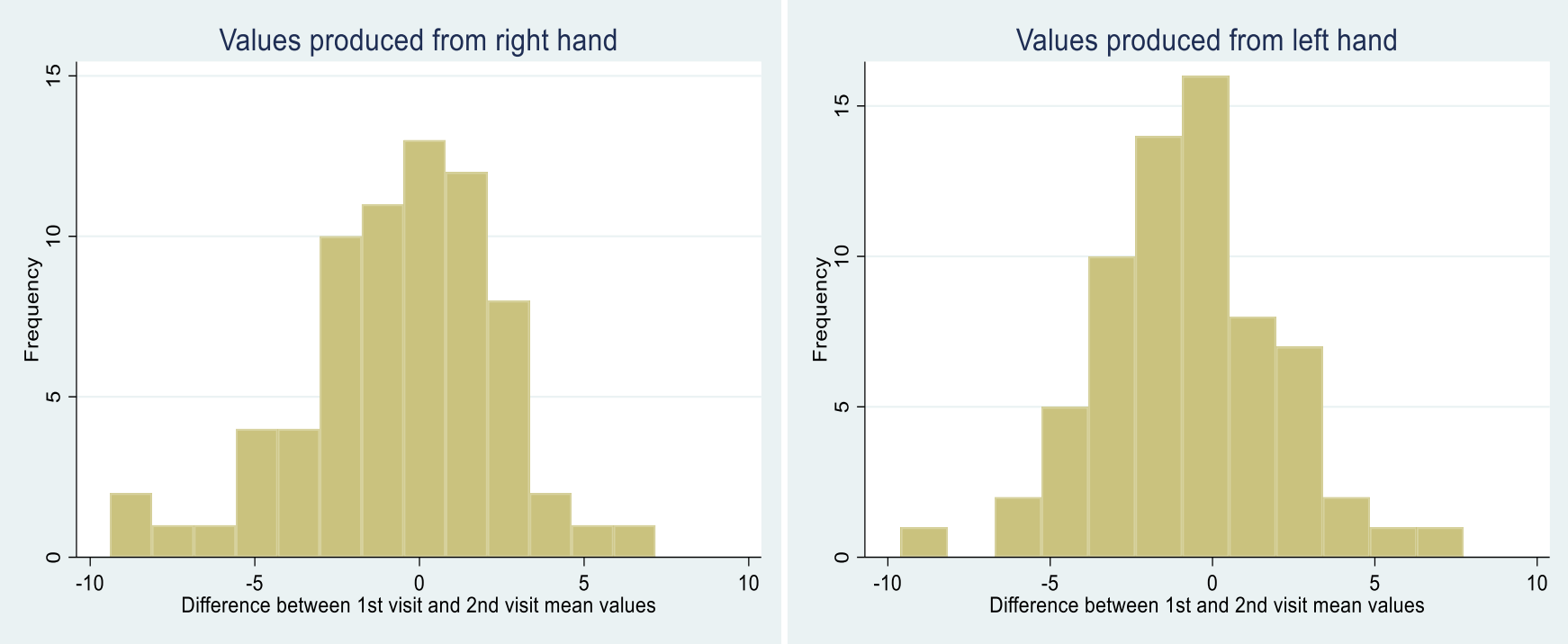
482. MacDermid, J.C., et al., *Interrater reliability of pinch and grip strength measurements in patients with cumulative trauma disorders.* Journal of Hand Therapy, 1994. **7**(1): p. 10-14.
483. Maher, C. and Y. Seoyoung, *Reliability of the Bulb Dynamometer for Assessing Grip Strength.* Open Journal of Occupational Therapy (OJOT), 2018. **6**(2): p. 1-7.
484. Mawdsley, R.H., et al., *Reliability of an alternative position of measuring grip strength in elderly females.* Issues on Aging, 2001. **24**(1): p. 7-10.
485. Medina-Mirapeix, F., et al., *Interobserver Reliability of Peripheral Muscle Strength Tests and Short Physical Performance Battery in Patients With Chronic Obstructive Pulmonary Disease: A Prospective Observational Study.* Archives of Physical Medicine & Rehabilitation, 2016. **97**(11): p. 2002-2005.
486. Niebuhr, B.R., R. Marion, and M.L. Fike, *Reliability of grip strength assessment with the computerized Jamar dynamometer.* Occupational Therapy Journal of Research, 1994. **14**(1): p. 3-18.
487. Paltamaa, J., et al., *Reliability of physical functioning measures in ambulatory subjects with MS [corrected] [published erratum appears in PHYSIOTHER RES INT 2006;11(2):123].* Physiotherapy Research International, 2005. **10**(2): p. 93-109.
488. Peolsson, A., R. Hedlund, and B. Öberg, *Intra- and inter-tester reliability and reference values for hand strength.* Journal of Rehabilitation Medicine (Taylor & Francis Ltd), 2001. **33**(1): p. 36-41.
489. Plant, C.E., et al., *A comparison of electronic and manual dynamometry and goniometry in patients with fracture of the distal radius and healthy participants.* Journal of Hand Therapy, 2016. **29**(1): p. 73-79.
490. Puthoff, M.L. and D. Saskowski, *Reliability and responsiveness of gait speed, five times sit to stand, and hand grip strength for patients in cardiac rehabilitation.* Cardiopulmonary Physical Therapy Journal, 2013. **24**(1): p. 31-7.
491. Savva, C., C. Karagiannis, and A. Rushton, *Test-retest reliability of grip strength measurement in full elbow extension to evaluate maximum grip strength.* The Journal of hand surgery, European volume, 2013. **38**(2): p. 183-186.
492. Schreuders, T.A.R., et al., *Measurement error in grip and pinch force measurements in patients with hand injuries.* Physical Therapy, 2003. **83**(9): p. 806-815.
493. Segura-Ortí, E. and F.J. MartÃ½nez-Olmos, *Test-Retest Reliability and Minimal Detectable Change Scores for Sit-to- Stand-to-Sit Tests, the Six-Minute Walk Test, the One-Leg Heel-Rise Test, and Handgrip Strength in People Undergoing Hemodialysis.* Physical Therapy, 2011. **91**(8): p. 1244-1252.
494. Shechtman, O., L. Gestewitz, and C. Kimble, *Reliability and validity of the DynEx dynamometer.* Journal of Hand Therapy, 2005. **18**(3): p. 339-347.
495. Silva, A.G., et al., *Inter-rater reliability, standard error of measurement and minimal detectable change of the 12-item WHODAS 2.0 and four performance tests in institutionalized ambulatory older adults.* Disability & Rehabilitation, 2019. **41**(3): p. 366-373.
496. Smidt, N., et al., *Interobserver reproducibility of the assessment of severity of complaints, grip strength, and pressure pain threshold in patients with lateral epicondylitis.* Archives of Physical Medicine & Rehabilitation, 2002. **83**(8): p. 1145-1150.
497. Solari, A., et al., *Reliability of clinical outcome measures in Charcot-Marie-Tooth disease.* Neuromuscular Disorders, 2008. **18**(1): p. 19-26.
498. Spijkerman, D.C.M., et al., *Standardization of grip strength measurements. Effects on repeatability and peak force.* Scandinavian Journal of Rehabilitation Medicine, 1991. **23**(4): p. 203-206.
499. Stephens, J.L., N. Pratt, and S. Michlovitz, *The reliability and validity of the Tekdyne hand dynamometer: part II.* Journal of Hand Therapy, 1996. **9**(1): p. 18-26.
500. Stockton, K.A., et al., *Test-retest reliability of hand-held dynamometry and functional tests in systemic lupus erythematosus.* Lupus, 2011. **20**(2): p. 144-150.

501. Svensson, E. and C. Häger-Ross, *Hand function in Charcot-Marie-Tooth: test-retest reliability of some measurements.* Clinical Rehabilitation, 2006. **20**(10): p. 896-908.
502. Tager, I.B., A. Swanson, and W.A. Satariano, *Reliability of physical performance and self-reported functional measures in an older population.* Journals of Gerontology - Series A Biological Sciences and Medical Sciences, 1998. **53**(4): p. M295-M300.
503. Tan, B., et al., *Grip strength measurement in competitive ten-pin bowlers.* Journal of Sports Medicine & Physical Fitness, 2001. **41**(1): p. 68-72.
504. Trippolini, M., et al., *Reliability and Safety of Functional Capacity Evaluation in Patients with Whiplash Associated Disorders.* Journal of Occupational Rehabilitation, 2013. **23**(3): p. 381-390.
505. Trutschnigg, B., et al., *Precision and reliability of strength (Jamar vs. Biodex handgrip) and body composition (dual-energy X-ray absorptiometry vs. bioimpedance analysis) measurements in advanced cancer patients.* Applied Physiology, Nutrition & Metabolism, 2008. **33**(6): p. 1232-1239.
506. Tsang, R.C.C., *Reference values for 6-Minute Walk Test and hand-grip strength in healthy Hong Kong Chinese adults.* Hong Kong Physiotherapy Journal, 2005. **23**: p. 6-12.
507. Tveter, A.T., et al., *Measuring Health-Related Physical Fitness in Physiotherapy Practice: Reliability, Validity, and Feasibility of Clinical Field Tests and a Patient-Reported Measure.* Journal of Orthopaedic & Sports Physical Therapy, 2014. **44**(3): p. 206-216.
508. Vermeulen, J., et al., *Measuring Grip Strength in Older Adults: Comparing the Grip-ball With the Jamar Dynamometer.* Journal of Geriatric Physical Therapy, 2015. **38**(3): p. 148-153.
509. Villafane, J.H., et al., *Reliability of handgrip strength test in elderly subjects with unilateral thumb carpometacarpal osteoarthritis.* Hand, 2015. **10**(2): p. 205-9.
510. Villafane, J.H., et al., *Reliability of the Handgrip Strength Test in Elderly Subjects With Parkinson Disease.* Hand, 2016. **11**(1): p. 54-8.

Appendices

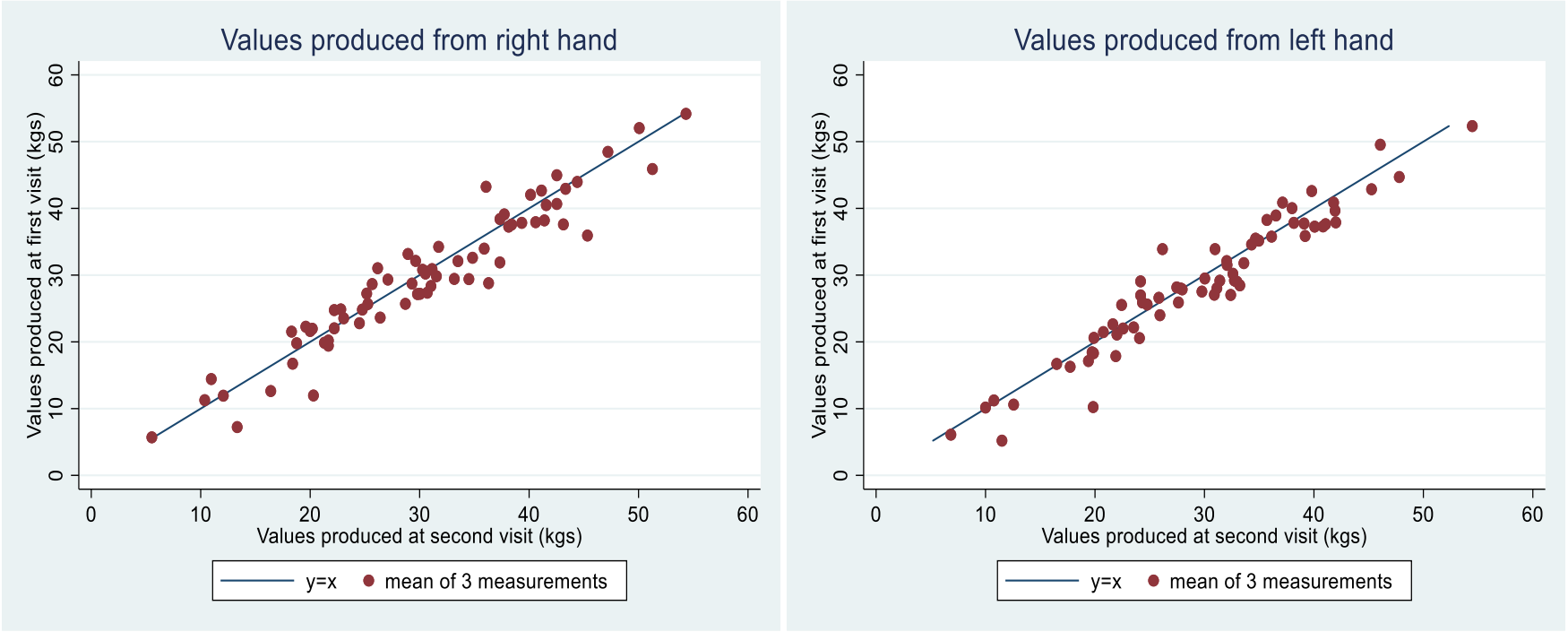
Appendix A: Additional figures for Chapter 3

Appendix A1. Histogram of the difference between the mean values produced at the first and second visit.

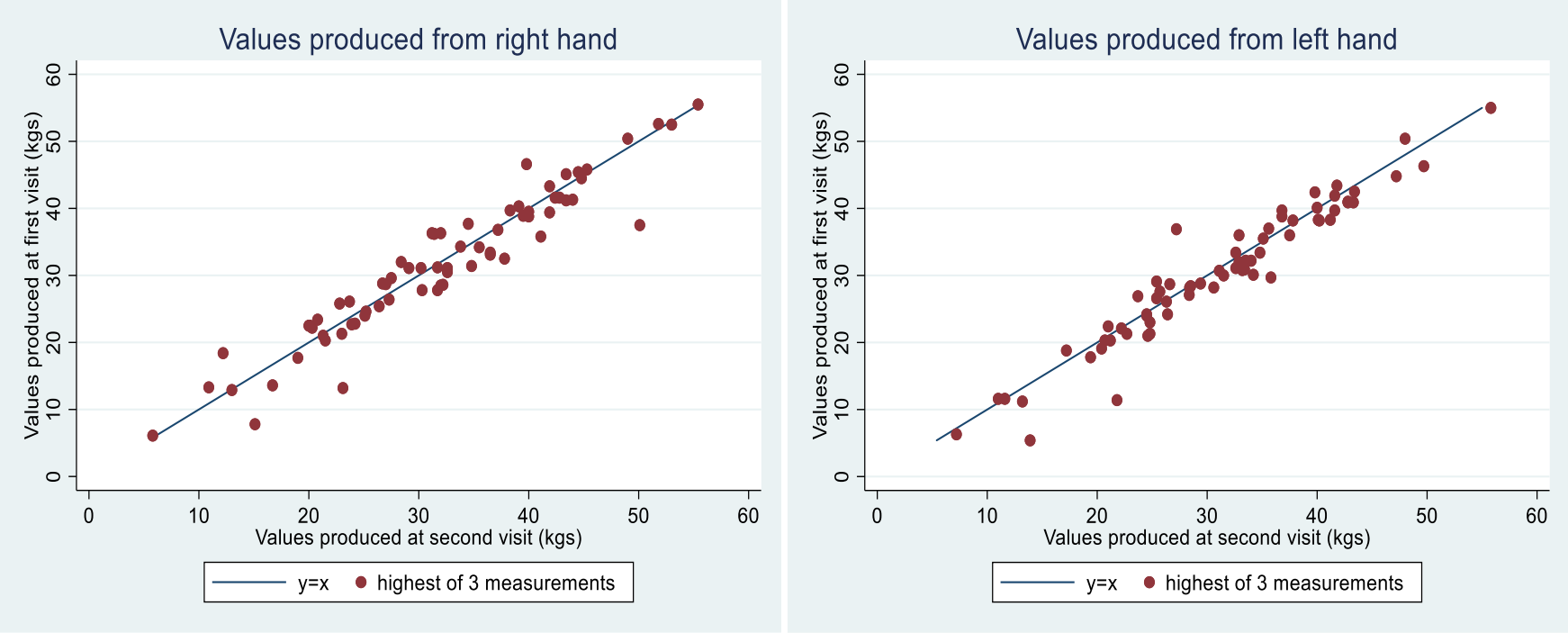


Appendix A2. Histogram of the difference between the highest values produced at the first and second visit.

Appendix A3. Scatterplot of the mean of 3 within-individual measurements produced at each visit.



Appendix A4. Scatterplot of the highest of 3 within-individual measurements produced at each visit.



Appendix B: Search strategy and additional tables for Chapter 4

Appendix B1. Search strategy.

The search strategy for MEDLINE was:

1 (variabil$ or reliabil$ or reproducibil$ or repeatabil$ or replicab$).m_titl.

2 ((measure$ or assess$ or test$ or assay$ or instrument$) and (properties or method$)).m_titl.

3 (measure$ or assess$ or test$ or assay$ or instrument$).m_titl. and (variabil$ or reliabil$ or repeatabil$ or precis$ or reproducibil$).ab.

4 (biol$ and (variat$ or variab$)).m_titl.

5 (biol$ adj3 (variat$ or variab$)).ab.

6 ((inter or intra) adj (rater or tester or technician or examiner or assay or individual or participant)).m_titl.

7 ((inter or intra) adj (tester or technician or examiner or assay or individual or participant)).ab. (18075)

8 (retest or re-test).m_titl.

9 COSMIN.ab.

10 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9

11 MEDLINE.tw.

12 systematic review/

13 systematic review.tw.

14 meta-analysis.pt.

15 (meta adj analy$).tw.

16 metaanaly$.tw.

17 11 or 12 or 13 or 14 or 15 or 16

18 (biography or case reports or comment or directory or editorial or festschrift or interview or lectures or legislation or letter or news or newspaper article or patient education handout).pt. (3668459)

19 exp Animals/

20 exp Humans/

21 19 not 20

22 18 or 21

23 10 and 17

24 23 not 22

25 limit 24 to yr="2010 –Current”

The search strategy for Embase was:

1 (variabil* or reliabil* or reproducibil* or repeatabil* or replicab*).ti. (106533)

2 ((measure* or assess* or test* or assay* or instrument*) and (properties or method*)).ti.

3 (measure* or assess* or test* or assay* or instrument*).ti. and (variabil* or reliabil* or repeatabil* or precis* or reproducibil*).ab.

4 (biol* and (variat* or variab*)).ti.

5 (biol* adj3 (variat* or variab*)).ab.

6 ((inter or intra) adj (rater or tester or technician or examiner or assay or individual or participant)).ti.

7 ((inter or intra) adj (tester or technician or examiner or assay or individual or participant)).ab.

8 (retest or re-test).ti.

9 COSMIN.ab.

10 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9

11 MEDLINE.ti,ab.

12 exp systematic review/

13 "systematic review".ti,ab.

14 meta-analysis/ (159302)

15 (meta adj analy*).ti,ab.

16 metaanaly*.ti,ab.

17 11 or 12 or 13 or 14 or 15 or 16

18 (biography or "case reports" or comment or directory or editorial or festschrift or interview or lectures or legislation or letter or news or "newspaper article" or "patient education handout").pt.

19 exp Animals/

20 exp Humans/

21 19 not 20

22 18 or 21

23 10 and 17

24 23 not 22

25 limit 24 to yr="2010 -Current"

Appendix B2. Data extraction form.

A: Review information	
A1	Title
A2	Medical area
A3	Number of primary studies included
A4	Number of tests examined
A3	Type of variability examined (e.g., inter/intra-observer, test-retest)
A4	Was test variability the main aim of the review?
B: Test information	
B1	Name of test
B2	Type of test (e.g., laboratory, imaging)
C: Searching of primary studies	
C1	Databases listed
C2	Was the search strategy presented?
C3	Was the search period clearly described?
D: Review methods	
D1	Were the criteria for inclusion clearly described?
D2	Was a PRISMA flow chart presented?
D3	Were the characteristics of the included primary studies clearly presented?
D4	What was the approach used for article selection?
D5	What was the approach used for data extraction?
E: Quality assessment of primary studies	
E1	Was the quality of primary studies assessed?
E2	If yes, what was the approach used for quality assessment?
E3	If yes, which tool was used for quality assessment?
F: Statistical methods	
F1	What statistical parameters were reported?
F2	Was a synthesis of the reported estimates performed?
F3	If performed, provide details on how this was done

Appendix B3. Characteristics of the identified systematic reviews.

Author	Year	Medical area	Type of variability examined	Number of studies	Test information			Statistical parameters reported	Quantitative synthesis performed
					Number of tests	*Group*	*Name*		
Aarsand et al [153]	2018	Lipids, Enzymes, Diabetes, Kidney	Test-retest	128	>10	Laboratory	Multiple measurands	Coefficient of variation	Yes
Abou El Hassan [223]	2018	Not specific	Test-retest	56	>10	Laboratory	Multiple measurands	Coefficient of variation	No
Abou [224]	2018	Spinal Cord Injury	Inter/intra-observer & test-retest	8	>10	Physical (device and non-device based)	Hand-held dynamometry, upper body sway, maximal balance range, coordinated stability, alternating reach test, seated reach test, t-shirt test, Functional Reach (FR), Reach Area (RA), Bilateral Reach (BR), modified Functional Reach Test (mFRT), Limits of stability (LOS), Sequential Weight Shifting (SWS).	Intra class correlation, Kappa coefficient	No
Adhia [225]	2013	Joint kinematics	Inter/intra-observer & test-retest	7	1	Physical (device based)	Electromagnetic tracking devices (EMTD)	Intra class correlation, Pearson correlation, Standard error of measurement	No

Ager [226]	2017	Shoulder proprioception	Inter/intra-observer	21	10	Physical (device based)	Dynamometer, Inclinometer, Laser pointer, Goniometer, CPM, Fabricated Lab, AMEDA, Motion analysis, Photo analysis, iPod touch.	Intra class correlation, Standard error of measurement, Smallest detectable change, Cronbach's alpha	No
Aguilar [227]	2017	Osteoarthritis	Inter/intra-observer	11	1	Imaging	MRI	Intra class correlation, Pearson correlation, Coefficient of variation	No
Alcázar [228]	2011	Gynaecological cancer	Inter/intra-observer	46	1	Imaging	3D Ultrasound	Not reported	No
Aloraini [229]	2015	Spasticity	Inter/intra-observer & test-retest	4	2	Physical (device based)	Electrophysiologic measures, Force/torque measurements	Intra class correlation, Kappa coefficient, Tau-b	No
Alreni [230]	2017	Neck pain	Inter/intra-observer & test-retest	2	1	Physical (non-device based)	Single Arm Military Press (SAMP) test	Intra class correlation	No
Ammann-Reiffer [231]	2014	Neuromuscular disorders	Inter/intra-observer & test-retest	20	>10	Physical (device and non-device based)	7.5-Meter Shuttle Run Test, 10×5-Meter Sprint Test, 10-Meter Fast Walk Test, 10-Meter Shuttle Run Test, Six-Minute Walk Test, Time Up and Down Stairs Test, Fast 1-Minute Walk	Intra class correlation, Spearman correlation, Kappa coefficient, Standard error of measurement, Smallest	No

							Test, Full turn, Functional Walking Test, Stopwatch, Photoelectric cells linked to electronic timer, GAITRite, Maximal Speed During Treadmill Walking Test, Timed “Up & Go” Test	detectable change, Mean difference	
Artero [232]	2011	Physical fitness	Inter/intra-observer & test-retest	19	>10	Physical (device and non-device based)	Cardiorespiratory, Musculoskeletal, Motor, Body composition tests	Intra class correlation, Percentage agreement, Kappa coefficient, Limits of agreement, Pearson correlation, Standard error of measurement, Smallest detectable change, Paired t-test, Root Mean Square Error, Coefficient of variation, Kruskal-Wallis test, Technical Error of Measurement,	No

								Pitman's test of correlated variances, Wilcoxon test	
Avouac [233]	2010	Pulmonary arterial hypertension secondary to systemic sclerosis	Test-retest	1	1	Physical (non-device based)	6-minute walk test	Pearson correlation	No
Balemans [234]	2013	Cerebral Palsy	Inter/intra-observer & test-retest	7	5	Physical (device and non-device based)	Aerobic and anaerobic tests	Intra class correlation, Spearman correlation, Pearson correlation, Standard error of measurement, Smallest detectable change, Limits of agreement	No
Balzer [235]	2017	Upper motor neuron lesions	Test-retest	2	2	Physical (device and non-device based)	Selective voluntary motor control measures	Not reported	No
Barrett [236]	2014	Thoracic kyphosis	Inter/intra-observer	26	>10	Physiologic	Methods of measuring thoracic kyphosis	Intra class correlation	No
Bartels [237]	2013	Chronic paediatric conditions	Test-retest	9	1	Physical (non-device based)	6-minute walk Test	Intra class correlation, Pearson correlation, Limits of agreement,	No

								Standard error of measurement, Smallest detectable change	
Basile [238]	2017	Osteopathy	Inter/intra-observer	17	1	Clinical examination	Osteopathic diagnostic palpatory tests	Intra class correlation, Kappa coefficient, Fisher's exact test	No
Beales [239]	2011	Abdominal aortic aneurysm	Inter/intra-observer	9	1	Imaging	Ultrasound	Limits of agreement, reproducibility coefficients using generalized estimating equation	No
Beaulieu [240]	2017	Human motor systems	Inter/intra-observer	34	1	Physiologic	Transcranial magnetic stimulation	Intra class correlation, Pearson correlation, Concordance coefficient, Coefficient of variation, Typical percentage error, Standard error of measurement, Smallest	

								detectable change	
Bellet [241]	2012	Cardiac rehabilitation	Test-retest	14	1	Physical (non-device based)	6-minute walk test	Intra class correlation, Relative change	No
Bennett [242]	2016	Health and fitness	Test-retest	2	1	Physical (device based)	Submaximal Step Test	Limits of agreement	No
Berger [243]	2014	Arthritis	Inter/intra-observer	4	1	Imaging	Radiographic imaging test	Kappa coefficient	No
Bergquist [244]	2019	Physical performance	Inter/intra-observer & test-retest	9	>10	Physical (device and non-device based)	Muscle and balance tests	Not reported	No
Bernard [245]	2015	Schizophrenia	Test-retest	1	1	Physical (non-device based)	6-minute walk test	Intra class correlation, Pearson correlation	No
Bianco [246]	2015	Physical fitness	Test-retest	100	>10	Physical (device and non-device based)	Multiple fitness tests	Pearson correlation, Mean difference	No
Bieniek [247]	2014	Physical fitness	Inter/intra-observer & test-retest	11	>10	Physical (device and non-device based)	Multiple fitness tests	Intra class correlation, Kappa coefficient, Percentage agreement	No
Bohannon [248]	2011	Muscle strength	Test-retest	10	1	Physical (non-device based)	Five-repetition sit-to-stand test	Intra class correlation	No
Bohannon [249]	2017	Muscle strength	Test-retest	17	1	Physical (device based)	Hand-held dynamometer	Intra class correlation, Standard error of	No

								measurement, Smallest detectable change, Technical Error of Measurement	
Bohannon [250]	2019	Physical fitness	Test-retest	10	1	Physical (non-device based)	Two-minute step test	Intra class correlation	No
Borotikar [251]	2017	Musculoskeletal	Inter/intra-observer	20	1	Imaging	MRI	Intra class correlation	No
Braga [252]	2010	Diabetes	Test-retest	9	1	Laboratory	Hemoglobin	Coefficient of variation	No
Braga [253]	2012	Cardiovascular	Test-retest	11	1	Laboratory	C-reactive protein	Coefficient of variation	No
Brink [254]	2011	Neuro-musculoskeletal	Inter/intra-observer & test-retest	6	1	Imaging	3D posture-measuring instruments	Intra class correlation, Pearson correlation, Standard error of measurement, Technical error of measurement, Dunnett comparison test	No
Burgess [255]	2016	Colorectal cancer	Inter/intra-observer & test-retest	21	5	Physical (device based)	Hand grip strength, hand-held dynamometry, isometric strength, manual muscle testing, and trunk flexion	Intra class correlation	No

							strength/lower extremity (LE) dynamometry		
Carlsson [256]	2013	Lower back pain	Inter/intra-observer	7	>10	Physical (non-device based)	Screening test for assessing lower back pain	Intra class correlation, Kappa coefficient	No
Carobene [257]	2013	Liver disease	Test-retest	33	3	Laboratory	Alanine aminotransferase, aspartate aminotransferase, and γ-glutamyl transferase measurements	Coefficient of variation, Reference change value	No
Cavaleri [154]	2017	Assessment of central nervous system	Not clear	4	1	Physical (device based)	Transcranial magnetic stimulation	Intra class correlation	Yes
Chaabene [258]	2018	Physical fitness	Inter/intra-observer & test-retest	20	>10	Physical (device and non-device based)	Multiple fitness tests	Intra class correlation, Standard error of measurement, Limits of agreement, Pearson correlation coefficient, Paired t-test	No
Chamorro [155]	2017	Muscle strength	Intra-observer	15	1	Physical (device based)	Dynamometer	Standard error of measurement, Limits of agreement	Yes
Cheung [259]	2010	Rheumatoid Arthritis	Inter/intra-observer	35	1	Imaging	Ultrasonography	Intra class correlation, Coefficient of	No

								variation, Kappa coefficient, Percentage agreement, Kendall's W	
Childs [260]	2014	Measurement of the liver	Inter/intra-observer	9	1	Imaging	Ultrasound	Percentage agreement, Absolute difference	No
Chiwaridzo [261]	2017	Physical performance	Inter/intra-observer & test-retest	14	>10	Physical (non-device based)	Multiple fitness tests	Intra class correlation, Standard error of measurement, Coefficient of variation, Pearson correlation, Limits of agreement, Typical error of measurement	No
Clark [262]	2017	Neurological disorders in children	Inter/intra-observer	11	3	Physical (device and non-device based)	Multiple neurological tests	Intra class correlation, Pearson correlation, Standard error of measurement, Smallest detectable change	No

Clark [263]	2018	Standing balance	Test-retest	12	1	Physical (device based)	Wii Balance Board	Intra class correlation, Coefficient of variation, Concordance correlation, Smallest detectable change, Typical error of measurement	No
Crowley [264]	2016	Cardiovascular	Inter/intra-observer	41	1	Imaging	Echocardiography	Intra class correlation, Limits of agreement, Coefficient of variation, Total deviation index	No
Cutolo [265]	2018	Systemic sclerosis	Inter/intra-observer	1	1	Imaging	Laser speckle contrast analysis	Intra class correlation	No
De Albuquerque [266]	2018	Cervical lordosis	Inter/intra-observer	2	1	Imaging	X-ray photogrammetry	Intra class correlation	No
De Guio [267]	2016	Cerebral small vessel disease	Not clear	29	1	Imaging	MRI	Intra class correlation, Coefficient of variation, reproducibility error (%), Standard error of measurement, Smallest detectable	No

								change, Absolute & relative difference	
De Langen [137]	2012	Cancer	Test-retest	5	1	Imaging	PET with glucose 18F-FDG	Intra class correlation	Yes
De Paula Lima [268]	2011	Transversus abdominis muscle activity	Inter/ intra-observer	4	1	Physiologic	Pressure biofeedback unit	Intra class correlation, Coefficient of variation, Limits of agreement, Pearson correlation, Wilcoxon test	No
De Valk [269]	2016	Malrotation of femoral and/or tibial component	Inter/intra-observer	12	1	Imaging	Computed tomography	Intra class correlation	No
Décary [270]	2016	Knee disorders	Inter/ intra-observer	33	>10	Physical (non-device based)	Multiple physical tests	Intra class correlation, Kappa coefficient	No
DeJong [271]	2017	Dermatology	Inter/intra-observer & test-retest	6	1	Imaging	Ultrasound elastography	Intra class correlation, Kappa coefficient, Percentage agreement	No
Dekkers [272]	2014	Cerebral palsy	Inter/intra-observer & test-retest	7	5	Physical (device based)	Upper extremity muscle strength tests	Intra class correlation, Pearson correlation, Paired t-test	No

Deng [273]	2018	Scar maturation	Inter/intra-observer & test-retest	7	3	Physical (device based) & Imaging	Vascularity measurement devices	Intra class correlation, Kappa coefficient	No
Denteneer [274]	2017	Low back pain	Inter/intra-observer	16	>10	Physical (non-device based)	Multiple physical tests	Intra class correlation, Kappa coefficient	No
Denteneer [275]	2018	Low back pain	Inter/intra-observer & test-retest	20	>10	Physical (non-device based)	Multiple physical tests	Intra class correlation, Kappa coefficient, Percentage agreement	No
D'hondt [276]	2017	Shoulder pain	Inter/intra-observer	40	>10	Physical (device and non-device based)	Multiple physical tests	Not reported	No
Dobson [277]	2012	Hip and groin pathology	Inter/intra-observer	12	>10	Physical (device and non-device based)	Multiple physical tests	Intra class correlation, Kappa coefficient, Standard error of measurement, Smallest detectable change, Cronbach's alpha	No
Dobson [278]	2012	Osteoarthritis	Inter/intra-observer & test-retest	16	>10	Physical (non-device based)	Multiple physical tests	ICC, correlation coefficient, Goodman & Kruskal's gamma,	No

								Standard error of measurement, Smallest detectable change	
Ekpo [279]	2012	Breast cancer	Inter/intra-observer	2	1	Imaging	Digital breast tomosynthesis	Pearson correlation	No
English [280]	2012	Musculoskeletal	Inter/intra-observer & test-retest	24	1	Imaging	B-mode ultrasound	Intra class correlation, Standard error of measurement, Coefficient of variation, Pearson correlation, Mean difference, G coefficient	No
Fan [281]	2016	Neurological disorders	Test-retest	12	2	Imaging	Emission tomography and arterial spin MRI	Coefficient of variation	No
Ferreira [282]	2017	Dental Medicine	Inter/intra-observer	5	1	Imaging	Tomographic digital models	Intra class correlation, Limits of agreement, Paired t-test	No
Field [283]	2013	Motor impairments in children	Inter/intra-observer & test-retest	4	>10	Physical (non-device based)	Multiple physical tests	Pearson correlation	No
Fisher [284]	2015	Prostate cancer	Inter/intra-observer & test-retest	36	3	Physical (device based)	Strength and muscular endurance outcome measures	Intra class correlation, Spearman correlation	No

Flamand [285]	2013	Neuromuscular disorders	Inter/intra-observer & test-retest	11	9	Physical (device and non-device based)	Multiple physical tests	Intra class correlation, Standard error of measurement, Coefficient of variation	No
Fonseca [286]	2018	Hand nerve injuries	Inter/intra-observer & test-retest	6	5	Physical (device and non-device based)	Multiple physical tests	Intra class correlation, Kappa coefficient, Pearson correlation, Cronbach's alpha	No
Fotheringham [287]	2015	COPD	Not clear	41	8	Physical (non-device based)	Multiple physical tests	Intra class correlation, Pearson correlation, Mean difference	No
Gebruers [288]	2010	Stroke	Test-retest	7	1	Physical (device based)	Accelerometer	Intra class correlation, Pearson correlation	No
Gengo e Silva [289]	2014	Peripheral arterial occlusive disease	Inter-observer	2	2	Physiological & Imaging	Oscillometric device & Doppler ultrasound	Not reported	No
Giraud [290]	2017	Cardiac output	Inter-observer	16	1	Physiological	Transpulmonary thermodilution	Coefficient of variation, Precision (%), Least significant change	No

Golden [291]	2011	Hypothalamic–pituitary–adrenal axis (HPA)	Inter-observer	6	3	Laboratory	HPA axis measures	Intra class correlation, Pearson correlation, Spearman correlation	No
Gonzalez Lao [156]	2019	Diabetes mellitus	Test-retest	47	9	Laboratory	Multiple lab test results	Coefficient of variation	Yes
Gonzalez-Suarez [292]	2018	Carpal tunnel syndrome	Inter/intra-observer	3	1	Imaging	Ultrasonography	Intra class correlation, Kappa coefficient, Pearson correlation	No
Guillaud [293]	2016	Cranial osteopathy	Inter/intra-observer	9	1	Physical (non-device based)	Clinical assessment	Intra class correlation, Pearson correlation, Limits of agreement, Kappa coefficient	No
Hafsteinsdóttir [294]	2014	Stroke	Inter/intra-observer	3	1	Physical (non-device based)	Timed Up and Go Test	Intra class correlation	No
Hernaez [295]	2011	Fatty Liver	Inter/intra-observer	22	1	Imaging	Ultrasonography	Intra class correlation, Kappa coefficient, Percentage agreement	No
Himuro [296]	2017	Cerebral palsy	Test-retest	6	3	Physical (non-device based)	Multiple fitness tests	Intra class correlation, Kappa coefficient, Standard error	No

								of measurement, Smallest detectable change, Limits of agreement	
Hoogervorst [297]	2017	Fractures of the clavicle	Inter/intra-observer	3	5	Imaging	Imaging techniques to measure shortening of the midshaft clavicle fracture	Intra class correlation, Kappa coefficient, Paired t-test, Smallest detectable change	No
Hulteen [298]	2015	Movement Skill Competency	Inter/intra-observer & test-retest	16	8	Physical (non-device based)	Multiple fitness tests	Intra class correlation, Kendall coefficient of concordance, Kappa coefficient, Percentage agreement, Coefficient of variation, Pearson correlation	No
Hunter [157]	2011	Osteoarthritis	Inter/intra-observer	84	1	Imaging	MRI	Intra class correlation, Kappa coefficient, Coefficient of variation	Yes
Janaudis-Ferreira [299]	2012	COPD	Test-retest	4	4	Physical (non-device based)	Multiple fitness tests	Intra class correlation,	No

								Pearson correlation	
Jaspers [300]	2019	Burn wounds	Inter/intra-observer	2	2	Imaging	Videomicroscopy, Dermoscopy	Kappa coefficient	No
Johnston [301]	2017	COPD	Test-retest	13	5	Physical (non-device based)	Multiple fitness tests	Intra class correlation, Limits of agreement, Standard error of measurement, Smallest detectable change, Mean difference	No
Jonsson [302]	2018	Neck pain	Inter/intra-observer	11	1	Clinical examination	Physical tests	Intra class correlation, Kappa coefficient	No
Jørgensen [303]	2016	Wound healing	Inter/intra-observer	13	2	Imaging	Area & volume measurement techniques	Intra class correlation, Pearson correlation, Standard error of measurement, Coefficient of variation, Pillai's trace	No
Kasehagen [304]	2018	Peripheral nervous system	Inter/intra-observer	18	1	Imaging	Ultrasound	Intra class correlation, Standard error of measurement, Smallest detectable	No

								change, Limits of agreement, Kappa coefficient, Dependability coefficient	
Kleijn [158]	2012	Left ventricular dyssynchrony assessment	Inter/intra-observer	20	1	Imaging	3D echocardiography	Intra class correlation, Limits of agreement	Yes
Klingels [305]	2010	Hemiplegic cerebral palsy	Inter/intra-observer & test-retest	19	8	Physical (non-device based)	Multiple functional tests	Intra class correlation, Kappa coefficient, Pearson correlation	No
Konieczka [306]	2017	Shoulder pain	Inter/intra-observer	18	2	Clinical examination & Imaging	Palpation & Ultrasound	Intra class correlation, Standard error of measurement, Smallest detectable change, Coefficient of variation	No
Kramer [138]	2018	Cancer	Test-retest	3	1	Imaging	F–FLT positron emission tomography	Intra class correlation, Smallest detectable change, R square	Yes
Kristensen [307]	2017	Muscle strength in post-stroke hemiplegia	Test-retest	9	1	Physical (non-device based)	Isokinetic dynamometry	Intra class correlation, Standard error	No

								of measurement	
Kroman [308]	2014	Knee osteoarthritis	Intra-observer & test-retest	15	>10	Physical (device based)	Multiple fitness tests	Intra class correlation, Standard error of measurement, Coefficient of variation	No
Kwah [309]	2013	Fascicle lengths and pennation in human skeletal muscles	Inter/intra-observer	36	1	Imaging	Ultrasound	Coefficient of multiple correlation, Intra class correlation, Standard error of measurement, Coefficient of variation, Pearson correlation, Kappa coefficient	No
Lagarde [310]	2016	Dysphagia	Inter/intra-observer	4	1	Physiological	Cervical auscultation	Intra class correlation, Kappa coefficient, First-order agreement coefficient	No
Lamers [311]	2014	Multiple Sclerosis	Inter/intra-observer & test-retest	8	8	Physical (device and non-device based)	Multiple tests for upper limb assessment	Intra class correlation, Standard error of measurement, Coefficient of	No

								variation, Pearson correlation, Spearmen correlation	
Lange [147]	2017	Shoulder pathology	Inter/intra-observer	18	>10	Physical (non-device based)	Multiple physical tests	Intra class correlation, Kappa coefficient	Yes
Lange [312]	2017	Shoulder pathology	Inter/intra-observer	15	>10	Physical (non-device based)	Multiple physical tests	Not reported	No
Larsen [313]	2014	Scapular dyskinesis	Inter/intra-observer	30	>10	Clinical examination	Clinical assessment	Intra class correlation, Standard error of measurement, Limits of agreement, Kappa coefficient	No
Le Flao [314]	2018	Head & neck injuries	Test-retest	2	5	Physical (non-device based)	Head/neck response tests	Intra class correlation, Standard error of measurement	No
Lemeunier [315]	2017	Neck pain	Inter/intra-observer	3	7	Clinical examination	Clinical assessment	Kappa coefficient	No
Lemeunier [316]	2018	Neck pain	Inter/intra-observer	20	>10	Clinical examination	Clinical assessment	Intra class correlation, Kappa coefficient	No
Lennon [317]	2015	Cerebral palsy	Test-retest	6	3	Physical (device based)	Multiple fitness tests	Intra class correlation, Pearson correlation,	No

								Standard error of measurement, Smallest detectable change	
Liang [318]	2012	Smoke exposure	Test-retest	13	2	Laboratory	Urinary nicotine equivalents and plasma cotinine	Intra class correlation, Coefficient of variation	No
Lima [319]	2018	Cystic fibrosis	Test-retest	4	1	Physical (non-device based)	6-minute walk test	Intra class correlation, Pearson correlation, Mean difference, Coefficient of variation	No
Lisboa [320]	2015	Dental Medicine	Inter/intra-observer	14	1	Imaging	3D Cone-Beam computed tomography	Intra class correlation, Pearson correlation coefficient, Paired t-test	No
Maaswinkel [321]	2016	Lower back pain	Test-retest	4	4	Physical (non-device based)	Trunk stabilization assessments	Intra class correlation	No
MacKay [322]	2018	Knee osteoarthritis	Inter/intra-observer & test-retest	58	1	Imaging	MRI	Intra class correlation, Kappa coefficient, Coefficient of variation (CV), RMSCV (Root-mean-square CV)	No

Mahaffey [323]	2016	Obesity	Test-retest	2	1	Physical (non-device based)	6-minute walk test	Intra class correlation, Coefficient of variation, Limits of agreement	No
Maricar [324]	2016	Knee osteoarthritis	Inter/intra-observer	8	5	Clinical examination	Tests assessing knee joint effusion	Intra class correlation, Kappa coefficient	No
Marshall [325]	2018	Metatarsus adductus in newborns	Inter/intra-observer	4	6	Imaging	Tools to identify and quantify metatarsus adductus	Intra class correlation	No
May [326]	2010	Shoulder pain	Inter/intra-observer	37	>10	Physical (non-device based)	Physical examination tests	Intra class correlation, Kappa coefficient	No
Mc Auliffe [327]	2017	Tendinopathy	Inter/intra-observer	22	1	Imaging	Ultrasound	Intra class correlation, Standard error of measurement, Coefficient of variation, Smallest detectable change, Limits of agreement, Pearson correlation, Real mean square difference, Percentage error	No

McCreesh [328]	2015	Tendinopathy	Inter/intra-observer	20	4	Imaging	Radiological methods for measuring acromiohumeral distance	Intra class correlation, Standard error of measurement, Smallest detectable change, Coefficient of variation, Maximum difference (range, SD)	No
McVerry [329]	2012	Leptomeningeal Collateral Flow	Inter/intra-observer	7	2	Imaging	Catheter angiography, Computed tomography	Intra class correlation, Kappa coefficient, Percentage agreement	No
Michiels [330]	2013	Neck pain	Inter/intra-observer	6	3	Physiologic & Imaging	Laserpointer, Ultrasound, Electromagnetic trackers	Intra class correlation	No
Mieritz [331]	2012	Low back pain	Inter/intra-observer	14	6	Imaging	3D regional lumbar motion measurement systems	Intra class correlation, Pearson correlation, Standard error of measurement, Coefficient of variation, Limits of agreement, Root mean square error,	No

								Cronbach's alpha, Coefficient of multiple correlation	
Mijnarends [332]	2013	Sarcopenia	Test-retest	34	>10	Physical measure (device and non-device based)	Tools for measuring muscle strength & physical performance	Intra class correlation, Standard error of measurement, Pearson correlation, Limits of agreement	No
Milani [333]	2014	Body position measurement in rehabilitation	Inter/intra-observer	15	>10	Physical measure (device based)	Smartphone applications	Intra class correlation	No
Milne [334]	2018	Cerebellar ataxia	Inter/intra-observer	8	>10	Physical measure (non-device based)	Gait assessment	Intra class correlation, Cronbach's alpha, Pearson correlation, Spearman correlation, Standard error of measurement, Smallest detectable change, Coefficient of variation, Method error	No

Mohan [335]	2019	Diaphragmatic mobility	Inter/intra-observer	4	2	Imaging	Ultrasound & Radiography	Intra class correlation, Coefficient of variation	No
Mohseni Bandpei [336]	2014	Paraspinal muscle fatigue	Test-retest	12	1	Physiologic	Electromyography	Intra class correlation	No
Moloney [337]	2012	Musculoskeletal and neuropathic pain	Inter/intra-observer & test-retest	21	1	Physiologic	Thermal quantitative sensory testing	Intra class correlation, Smallest detectable change, Limits of agreement, Coefficient of variation	No
Moore [338]	2017	Community-dwelling older adults, Parkinson's disease, Huntington's disease, multiple sclerosis, vestibular disorders, post stroke, post unilateral transtibial amputation, knee pain and hip osteoarthritis	Inter/intra-observer & test-retest	12	1	Physical (non-device based)	Four square step test	Intra class correlation	No
Mulder-Brouwer [339]	2016	Cerebral palsy	Inter/intra-observer	7	1	Physical (device based)	Hand-held dynamometry	Intra class correlation, Standard error of	No

								measurement, Smallest detectable change, Limits of agreement, Coefficient of variation	
Muntaner-Mas [340]	2019	Physical fitness	Test-retest	2	2	Physical (device based)	Fitness apps	Intra class correlation, Coefficient of variation	No
Nae [341]	2017	Postural orientation error	Inter/intra-observer	25	10	Physical (non-device based)	Visual assessments and ratings	Intra class correlation, Standard error of measurement, Smallest detectable change, Kappa coefficient, First order agreement coefficient, Percentage agreement	No
Navarro [159]	2019	Adolescent Idiopathic Scoliosis	Inter/intra-observer	3	1	Imaging	Computed tomography	Intra class correlation	Yes
Neto [342]	2015	Obstetrics and gynaecology	Inter/intra-observer	112	1	Imaging	Ultrasound	Intra class correlation, Kappa coefficient	No
Nijholt [343]	2017	Muscle loss	Inter/intra-observer	13	1	Imaging	Ultrasound	Intra class correlation, Standard error of	No

								measurement, Limits of agreement	
Oberwahrenbrock [344]	2015	Multiple sclerosis	Intra-observer	27	7	Imaging	Optical coherence tomography	Intra class correlation, Limits of agreement	No
O'Meara [345]	2012	Pressure ulcers	Inter/intra-observer	8	3	Imaging	Wound measurement instruments	Intra class correlation, Standard error of measurement, Pearson correlation, Coefficient of variation, Mean difference	No
Ornetti [346]	2010	Osteoarthritis	Test-retest	3	1	Physical (non-device based)	Gait analysis	Intra class correlation	No
Ortega [347]	2015	Physical fitness	Inter/intra-observer & test-retest	21	3	Physical (device and non-device based)	Fitness tests	Intra class correlation, Standard error of measurement, Smallest detectable change, Cronbach's alpha, Pearson correlation, Paired t-test, Limits of agreement, Coefficient of	No

								variation, Kappa coefficient, Percentage agreement	
Özay [348]	2019	Olfaction	Test-retest	1	1	Physiologic	Retro-nasal test	Pearson correlation	No
Paech [349]	2011	Oncology	Inter-observer	6	1	Laboratory	Histology	Kappa coefficient	No
Pareira [350]	2014	Multiple (e.g., cardiac, COPD, cancer, cystic fibrosis, asthma)	Test-retest	18	1	Physical (non-device based)	Incremental shuttle walk	Intra class correlation, Pearson correlation, Coefficient of variation	No
Parry [351]	2015	Physiotherapy, Critical care	Inter/intra-observer & test-retest	47	>10	Imaging & Physical (device and non-device based)	Multiple tests for muscle mass, muscle strength and muscle function	Intra class correlation	No
Paul [352]	2016	Sports science	Inter-observer	21	3	Physical (non-device based)	Agility tests (multiple: use of stimuli (light, video, or human) to induce whole body change in velocity and/or direction	Intra class correlation, Coefficient of variation	No
Peeling [353]	2015	Infections disease (HIV)	Test-retest	32	1	Laboratory	CD4 enumeration (multiple assays)	Coefficient of variation	No
Perdomo [354]	2014	Breast cancer	Inter/intra-observer	51	6	Physiologic	Multiple (circumferential measurement, water displacement, bioelectrical impedance	Intra class correlation, Standard error of measurement, Kappa coefficient	No

							spectroscopy, perimetry, tonometry, and self-report)		
Petitclerc [355]	2015	Genetic disorder (myotonic dystrophy)	Inter/intra-observer	3	2	Physical (device and non-device-based)	Manual (MMT) or quantitative (QMT) muscle testing	Intra class correlation	No
Phillips [356]	2018	Emergency medicine	Inter-observer	2	1	Physiologic	Automated pupillometry	Not clear	No
Pin [357]	2014	Multiple (including healthy individuals)	Inter/intra-observer & test-retest	25	1	Physical (non-device based)	2-minute walk test	Intra class correlation, Pearson correlation coefficient, Limits of agreement, Smallest detectable change	No
Pollock [358]	2011	CVD (stroke)	Inter/intra-observer & test-retest	24	9	Physical (non-device based)	Walking balance (includes 4 single task measures; Step Test, Side-step Test and Four-square Step Test, Timed Up and Go)	Intra class correlation	No
Ponce-Garcia [359]	2018	Orthodontics	Inter-observer & test-retest	6	3	Imaging	Superimposition of 3D digital records obtained through Cone-Beam computed tomography (voxel based, landmark based, and surface based)	Intra class correlation	No

Pons [360]	2013	Cerebral palsy	Inter/intra-observer	19	3	Imaging	Imaging-based measures of hip geometry in children with cerebral palsy: X-ray, Computed tomography, Ultrasound	Intra class correlation, Standard error of measurement, Pearson correlation, Spearman correlation	No
Pons [361]	2018	Multiple	Inter/intra-observer & test-retest	30	1	Imaging	MRI (for skeletal muscle volume and shape); includes partially or completely automatic and manual techniques (slice-by-slice cross-sectional area (CSA) Segmentation; CSA segmentation on a reduced number of slices; CSA segmentation/ muscle thickness using a single slice and muscle length; CSA Segmentation on a single slice deformation of a parametric specific object (DPSO); deformation of a parametric specific object (DPSO), reduced MRI set method; other	Intra class correlation, Coefficient of variation	No

							automatic methods)		
Powden [139]	2015	Physiotherapy	Inter/intra-observer	12	1	Physical (non-device based)	Weight-bearing lunge test	Intra class correlation	Yes
Proud [362]	2015	Neurology (Parkinson's)	Intra-observer & test-retest	18	4	Physical (device and non-device-based)	Multiple measures of upper limb function (e.g., peg board tests and other timed tests)	Intra class correlation, Kappa coefficient, Smallest detectable change	No
Prowse [363]	2016	Musculoskeletal / Physiotherapy	Intra/inter-observer	14	4	Physical (device-based)	Postural asymmetry measures: scoliometer, iPhone, photography-based trunk aesthetic clinical evaluation tool (TRACE) and aesthetic index (AI), and assessment using plumb lines	Intra class correlation, Pearson correlation, Kappa coefficient, Mean difference	No
Rabelo [144]	2016	CVD	Test-retest	8	1	Physical (device-based)	Muscle strength assessment using dynamometers (computerized isokinetic dynamometers such as LIDO, Cybex II, Cybex 6000, Kin-Com, and Biodex System 3 Pro, or static multi-axial dynamometers)	Intra class correlation, Standard error of measurement	Yes

Ratter [364]	2014	Physiotherapy	Test-retest	14	10	Physical (non-device based)	Submaximal exercise tests: Astrand test; modified Astrand test; Lean body mass-based Astrand test; submaximal bicycle ergometer test following another protocol other than Astrand test; 2-km walk test; 5-minute, 6-minute and 10-minute walk tests; shuttle walk test; and modified symptom-limited Bruce treadmill test.	Intra class correlation, Cronbach's alpha, Limits of agreements	No
Reavis [149]	2015	Ears, nose, and throat	Test-retest	10	1	Physiologic	Distortion product Otoacoustic emission for cochlear function	Standard error of measurement	Yes
Reichmann [160]	2011	Rheumatology	Inter/intra-observer	24	1	Imaging	X-ray +/- Fluoroscopy to measure knee joint space	Intra class correlation, Coefficient of variation	Yes
Reis Durao [365]	2017	Maxillofacial	Inter/intra-observer	23	1	Imaging	Ultrasound for measuring muscle mass	Intra class correlation, Pearson correlation	No
Ringshausen [366]	2012	Infectious diseases	Test-retest	20	2	Laboratory	Interferon-gamma release assays: QuantiFERON-TB Gold or In-Tube	Intra class correlation	No

							version (QFT) and the T-SPOT.TB (T-SPOT)		
Roberts [367]	2011	Geriatrics	Inter-observer & test–retest	3	1	Physical (device based)	Dynamometers for measuring grip strength (Jamar, BTE Work Simulator, Martin Vigorimeter, Harpenden)	Pearson correlation	No
Robertson [368]	2014	Sport	Inter-observer & test-retest	22	>10	Physical (non-device based)	Skill outcomes in sport	Intra class correlation, Pearson correlation, Coefficient of variation, Limits of agreement, Typical error of measurement, G coefficient	No
Roeing [369]	2017	Geriatrics	Inter-observer & test-retest	3	5	Physical (device based)	Mobile phone assessment of balance and fall risk; static balance, TUG, Berg Balance Scale, 30s chair test, sit to stand test	Intra class correlation	No
Rondoni [370]	2017	Physical therapy	Intra/inter-observer	9	7	Physical (device based)	Active cervical range of motion measures (ACROM): universal inclinometer, standard dual-arm	Intra class correlation, Standard error of measurement,	Yes

							goniometer, gravity inclinometer, and cervical range of motion device, plus the Cybex Electronic Digital Inclinometer 320 (EDI-320), the Orthopaedic Systems Incorporated (OSI) Computerized Anatometry 6000 Spine Motion Analyzer (SMA), and the Flock-of-Birds system	Limits of agreement	
Rossini [371]	2016	Orthodontics	Intra/inter-observer	16	1	Imaging	Virtual dental study models	Intra class correlation, Pearson correlation, Houston coefficient, Cronbach's alpha, McNemar's test	No
Rozema [151]	2014	Ophthalmology	Inter/intra-observer	124	1	Physical (device based)	Biometric devices (e.g., Oculus Pentacam, Bausch & Lomb Orbscan, and Zeiss IOL Master)	Standard error of measurement	Yes
Rubio-Ochoa [372]	2016	Physical therapy	Inter-observer	5	3	Physical (non-device based)	Manual examination, cervical flexion-	Intra class correlation,	No

							rotation test (CFRT), combination	Kappa coefficient	
Ruhe [373]	2010	Physical therapy	Test-retest	32	1	Physical (device based)	Centre of pressure (COP) as a measure of postural stability	Intra class correlation, Coefficient of variation, G coefficient, Pearson correlation	No
Rydwik [374]	2012	Physical therapy, Geriatrics	Test-retest	3	2	Physical (non-device based)	Walking speed measurement (habitual or maximal) over varying distances	Intra class correlation, Pearson correlation	No
Saccomanno [375]	2015	Orthopaedics	Inter/intra-observer	11	1	Imaging	MRI	Intra class correlation, Kappa coefficient, Pearson correlation	No
Saether [376]	2013	Paediatrics (cerebral palsy)	intra/inter-observer & test–retest	35	>10	Physical (non-device based)	Physical tests to assess balance (e.g., Timed up and Go, and Timed Up and Down Stairs)	Intra class correlation, Standard error of measurement, Spearman correlation, G coefficient, Kappa coefficient	No
Salamh [161]	2019	Musculoskeletal	Inter/intra-observer	18	7	Physical (device and non-device based)	Low flexion, extension with internal rotation, HA, internal rotation, diagnostic	Intra class correlation, Standard error of measurement	Yes

							ultrasound, scapular-plane adduction, and myotonometer		
Saltzherr [377]	2014	Rheumatology	Inter/intra-observer	25	6	Imaging	Computed tomography (CT), Ultrasound, MRI, Positron emission tomography (PET), Single-photon emission CT and Scintigraphy	Intra class correlation, Kappa coefficient	No
Sam [378]	2019	Dentistry	Inter/intra-observer	13	1	Imaging	3D cephalometric landmarks in Cone-Beam computed tomography	Intra class correlation, Pearson correlation	No
Scalco [379]	2018	Paediatrics (cardio-respiratory)	Test-retest	11	4	Physical (non-device based)	Tests for functional capacity (e.g., 6-minute walk test)	Intra class correlation, Pearson correlation	No
Scheetz [380]	2015	Ophthalmology	Inter/intra-observer	32	4	Physical (device based)	Multiple glaucoma assessments: optic disc photographs, visual fields, ophthalmoscopy, combination.	Kappa coefficient	No
Schrama [381]	2014	Physical therapy	Intra-observer	54	1	Physical (device based)	Hand-held dynamometer for upper extremity assessment	Intra class correlation, Pearson correlation	No
Seagar [382]	2019	Musculoskeletal	Inter/intra-observer	5	6	Physical (device and non-device based)	Multiple instruments for assessment of cervical spine: goniometer, protractor, ROM	Intra class correlation, Standard error of measurement, Kappa	No

							limitation scale, Muscle Function Scale	coefficient, Pearson correlation	
Sepriano [383]	2015	Rheumatology	Intra-observer	8	1	Imaging	Dual-energy X-ray absorptiometry	Intra class correlation, Coefficient of variation	No
Serai [152]	2017	Liver disease	Test-retest	12	1	Imaging	MR elastography	Smallest detectable change (estimated using the coefficient of variation rather than the standard error of measurement)	Yes
Shiel [384]	2018	Sport science	Intra-observer	11	1	Physical (device based)	Dual energy X-ray absorptiometry for measuring body composition	Intra class correlation, Pearson correlation, Limits of agreement, Standard error of measurement, technical error of measurement, Percentage change in mean	No
Silva [385]	2014	Neurology	Intra-observer & test-retest	11	2	Physical (non-device based)	Sit to stand/stand to sit tests	Intra class correlation, Standard error	No

								of measurement, Smallest detectable change	
Simperingham [386]	2016	Physical therapy	Test-retest	34	4	Physical (device based)	Sprint acceleration profiling; radar and laser technology, and (2) non-motorised treadmill (NMT) and torque treadmill (TT) technology	Intra class correlation, Pearson correlation, Coefficient of variation, Limits of agreement, Standard error of measurement	No
Singh [387]	2014	Chronic respiratory disease	Test-retest	40	3	Physical (non-device based)	Field walking tests	Intra class correlation, Coefficient of variation, Limits of agreement	No
Sman [388]	2013	Musculoskeletal	Inter/intra-observer	7	8	Clinical examination	8 different clinical tests	Intra class correlation, Percentage agreement	No
Smith [389]	2013	Musculoskeletal	Inter/intra-observer	18	4	Physical (device and non-device based)	Tests for knee joint position sense (JPS)	Coefficient of variation, Pearson correlation	No
Smith [390]	2011	Orthopaedics	Inter/intra-observer	11	4	Imaging	Patellar instability; X-Ray, Computed tomography, MRI, Ultrasound	Intra class correlation, Mean difference in measurement	No

								between observers (SD)	
Sollis [391]	2014	Infectious diseases	Test-retest	37	6	Laboratory	HIV Viral Load (multiple assays)	Not reported	No
Southerest [392]	2013	Musculoskeletal	Inter/intra observer & Test-retest	9	2	Physical (non-device based)	Body pain diagrams	Intra class correlation	No
Stephenson [393]	2014	Paediatrics (haemophilia)	Test-retest	41	9	Physical (device and non-device based)	Multiple functional assessments in children	Intra class correlation	No
Stienen [394]	2019	Multiple spinal disease	Test-retest	82	>10	Physical (non-device based)	Multiple functional impairment tests (e.g., timed up and go, 6-minute walk test)	Intra class correlation	No
Stovall [395]	2010	Asymmetry in the lumbar spine and pelvis	Inter/intra-observer	7	1	Clinical examination	Palpation	Kappa coefficient	No
Symonds [396]	2017	Physical therapy	Test-retest	Unclear	5	Physical (non-device and device-based)	Multiple muscle and performance-based tests	Intra class correlation, Pearson correlation	No
Tagmouti [28]	2014	Infectious diseases	Test-retest	26	2	Laboratory	Interferon gamma (IFN-g) release assays (QuantiFERON and T-SPOT.TB)	Intra class correlation, Pearson correlation, Coefficient of variation	Yes
Talma [397]	2013	Obesity	Test-retest	20	>10	Physical (device-based)	Multiple bioelectrical impedance analysis	Intra class correlation, Pearson	No

							to estimate body fat %	correlation, Standard error of measurement, coefficient of variation, mean difference	
Tarara [398]	2016	Physical therapy	Test-retest	11	6	Physical (non-device based)	Multiple physical performance tests: closed kinetic chain upper extremity stability test (CKCUEST), seated shot put (2 hands), unilateral seated shot put, medicine ball throw, modified push-up test and 1-arm hop test	Intra class correlation, Pearson correlation	No
Terwee [399]	2011	Physical therapy	Test-retest	1	1	Physical (device-based)	Pedometer	Pearson correlation	No
Timmer [400]	2018	Haemophilia	Test-retest	14	3	Physical (device and non-device based)	Accelerometer, 6MWT, timed up and go test	Limits of agreement, standard error of measurement	No
Tong [401]	2018	Dry eyes	Inter-observer	3	1	Imaging	Tear scope	Kappa coefficient, paired t-test	No
Toohey [402]	2015	Physical therapy	Test-retest	5	1	Physical (device-based)	Sphygmo-manometer for muscle testing	Intra class correlation, Standard error of	No

								measurement, Coefficient of variation	
Traverso [403]	2018	Oncology	Test-retest	41	4	Imaging	Computed tomography (CT), Cone-Beam CT, Positron emission tomography (PET), MRI	Intra class correlation, concordance correlation, Spearman correlation	No
Valet [404]	2017	Multiple sclerosis	Test-retest	48	6	Physical (non-device and device-based)	Whole-body maximal exercise testing with gas exchange, whole-body maximal exercise testing without gas exchange analysis (cycle ergometer), whole-body submaximal exercise testing with gas exchange analysis (cycle ergometer), 6-minute walk test, Modified Canadian Aerobic Fitness Test (mCAFT), Ruffier-Dickson Test	Intra class correlation	No
van Bloemendaal [405]	2012	Cardiovascular (stroke)	Test-retest	32	>10	Physical (non-device and device-based), Imaging	Walking distance (2MWT, 6MWT, 12MWT), walking speed (various distances of comfortable or fast walk tests,	Intra class correlation, Kappa coefficient, Pearson correlation	No

							footswitch system, accelerometer, pedometer), functional ambulation (6MWT in different environments, functional ambulation categories, dynamic gait index, functional gait assessment), walking on different surfaces (6MWT on parquet and carpet)		
van de Pol [406]	2010	Musculoskeletal	Inter-observer	21	5	Physical (non-device and device-based), Imaging	Passive physiological or accessory movement of upper extremity joints (inclinonmeter, goniometer, visual/manual, tape measure, pollexograph)	Intra class correlation, Kappa coefficient, Pearson correlation	No
van de Water [407]	2016	Musculoskeletal	Inter/intra-observer & test-retest	13	8	Physical (non-device and device-based), Imaging	Diastasis of the rectus abdominis muscle measurement, 'finger width' method, tape measure, callipers, ultrasound, MRI, CT	Intra class correlation, Concordance correlation, (weighted) Kappa coefficient, Standard error	No

							and intraoperative measurements with a ruler or surgical compass	of measurement, Smallest detectable change, Limits of agreement	
van Kooij [408]	2017	Physiotherapy/ Physical medicine	Test-retest	15	1	Physical (device-based)	Goniometry (protractor-based)	Intra class correlation, Spearman correlation, Pearson correlation, Standard error of measurement, Smallest detectable change, Percentage agreement	No
van Trijffel [409]	2010	Musculoskeletal	Inter-observer	17	4	Physical (non-device and device-based), Imaging	Passive physiological or accessory movement of lower joints (goniometer, inclinonmeter, plurimeter, visual/manual)	Intra class correlation, Kappa coefficient, Pearson correlation	No
Wang [410]	2014	Rheumatology (Pediatric osteoporosis)	Test-retest	21	1	Imaging	Quantitative ultrasound (QUS)	Coefficient of variation	No
Weiner [29]	2017	Pediatric - cardiovascular	Test-retest	49	1	Physiologic	Heart rate variability	Intra class correlation, Pearson correlation, Spearman	Yes

								correlation (transformed using Fisher's Z)	
Welton [141]	2015	Neurology	Test-retest	23	5	Imaging	Graph-theoretic brain network metrics measured by functional MRI (fMRI), diffusion tensor imaging (DTI), magnetoencephalography, functional near-infrared spectroscopy (fNIRS) and arterial spin labelling	Intra class correlation	Yes
Wen [411]	2018	Sports science	Test-retest	6	1	Physical (non-device)	Loughborough soccer passing test	Intra class correlation, Pearson correlation, Coefficient of variation, Limits of agreement, Standard error of measurement	No
Wetterslev [412]	2016	Critical care	Not clear	2	1	Imaging	Echocardiography	Precision (%)	No
Williams [413]	2010	Physiotherapy	Inter/intra-observer & test-retest	46	>10	Physical (non-device and device-based); Imaging	Digital inclinometry, Electromagnetic motion analysis, Goniometry, Gravity-plus-compass	Intra class correlation, Kappa coefficient	No

							goniometry, Inclinometry, Optical motion, potentiometry, Tape measure, Ultrasound, Visual estimation, Miscellaneous		
Winser [414]	2015	Neurology	Test-retest	21	>10	Physical (nature of tests difficult to determine)	Multiple physical tests	Intra class correlation, Standard error of measurement, Smallest detectable change	No
Wouters [415]	2017	Neurology (Developmental disability)	Test-retest	26	>10	Physical (non-device and device-based)	Body composition (BIA, BMI, skinfold measurements, waist circumference), muscular strength (grip strength, hand-held dynamometry, softball throw, standing long jump), muscular endurance (arm hang, bench press, dumbbell press, isometric push-up, pull-up, sit-up) and cardiorespiratory fitness (fixed distance run/walk,	Intra class correlation, Kappa coefficient, Standard error of measurement, Limits of agreement	No

							fixed time run/walk, SRT, step test).		
Yang [416]	2017	Neurology - multiple conditions	Test-retest	23	>10	Physical (non-device based)	Multiple dual task balance and walking test	Intra class correlation, Standard error of measurement, Smallest detectable change	No
Yang [417]	2015	Geriatrics	Test-retest	26	>10	Physical (non-device based)	Multiple dual task balance and walking tests	Intra class correlation, Standard error of measurement, Smallest detectable change	No
Yoon [30]	2016	Oncology	Inter/intra-observer	9	1	Imaging	Computed tomography (tumour burden measurement using RECIST)	Relative measurement difference, Limits of agreement	Yes
Zanudin [418]	2017	Cerebral palsy	Inter/intra-observer & test-retest	16	6	Physical (non-device based)	Timed up and down stairs, timed up and go test, 1-minute walk test, functional walk test, 10-meter fast walk test, 6-minute walk test	Intra class correlation, Limits of agreement, Coefficient of variation, Standard error of measurement	No
Zayat [419]	2016	Rheumatology	Inter/intra-observer	5	1	Imaging	Ultrasound	Kappa coefficient	No

Zimmerman [420]	2017	Orthodontics	Inter/intra-observer	42	1	Imaging	Cone beam computed tomography	Intra class correlation, Pearson correlation, Dahlberg formula	No

Appendix B4. Details on quality items for each systematic review.

Author	Inclusion criteria specified	Study details presented	Databases listed	Search terms described	Search period described	Approach to screening	PRISMA flow chart	Approach to data extraction	Approach to quality assessment	Quality assessment tool(s) used
Aarsand et al [153]	Yes	No	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Not reported	No	Not reported	Independent and duplicate	BIVAC [153]
Abou El Hassan [223]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	No	Not reported	No	Not reported	Not reported	Author's own
Abou [224]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]
Adhia [225]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	No	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS [422] and QAREL [423]
Ager [226]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	COSMIN [421] and QualSyst [424]
Aguilar [227]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Alcázar [228]	Yes	Yes (tabulated)	MEDLINE only	Partially (terms in text only)	Yes	Single reviewer	No	Not reported	Not conducted	N/A
Aloraini [229]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not reported	Law and MacDermid [425]
Alreni [230]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]

Ammann-Reiffer [231]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [426]
Artero [232]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	No	Not reported	Independent and duplicate	Tool used in Essendrop et al [427]
Avouac [233]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Not reported	Jadad criteria [428]
Balemans [234]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not reported	COSMIN [134]
Balzer [235]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	No	Not reported	Independent and duplicate	COSMIN [134]
Barrett [236]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Yes	Single reviewer with checks by a second	Yes	Single reviewer	Independent and duplicate	Brink and Louw [429]
Bartels [237]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]
Basile [238]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Independent and duplicate	QAREL [423]
Beales [239]	Yes	Yes (tabulated)	MEDLINE only	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Not clear
Beaulieu [240]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Single reviewer	Yes	Single reviewer	Independent and duplicate	Law and MacDermid [425] and Chipchase et al [430]
Bellet [241]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Single reviewer	Yes	Single reviewer	Independent and duplicate	Brink and Louw [429]

						with checks by a second		with checks by a second		
Bennett [242]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Independent and duplicate	Not reported	Quality assessment tool for quantitative studies [available at INSTRUCTIONS FOR COMPLETION: Please circle appropriate response in each section (ephpp.ca)]
Berger [243]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Not clear
Bergquist [244]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Single reviewer	Yes	Not reported	Independent and duplicate	COSMIN [134]
Bernard [245]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Single reviewer	No	Independent and duplicate	Not conducted	N/A
Bianco [246]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	No	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Bieniek [247]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Bohannon [248]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Single reviewer	No	Single reviewer	Not conducted	N/A
Bohannon [249]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Partially (only start	Single reviewer	No	Single reviewer	Single reviewer	COSMIN [421]

					or end date given)					
Bohannon [250]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Not reported	Yes	Not reported	Not reported	Author's own
Borotikar [251]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Author's own
Braga [252]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Yes	Not reported	No	Not reported	Not conducted	N/A
Braga [253]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Yes	Not reported	No	Not reported	Not conducted	N/A
Brink [254]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Not reported	Not reported	N/A
Burgess [255]	No	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Cancer EDGE Task Force Rating Form [431] (Appendix A)
Carlsson [256]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Single reviewer	Independent and duplicate	QAREL [423]
Carobene [257]	No	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Yes	Not reported	No	Not reported	Not conducted	N/A
Cavaleri [154]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423], tool from Bialocerkowski et al [432], Chipchase et al [430]
Chaabene [258]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Tool used in Robertson et al [433]

Chamorro [155]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not reported	COSMIN [421]
Cheung [259]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Single reviewer	Yes	Not reported	Not reported	Author's own
Childs [260]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	Author's own
Chiwaridzo [261]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Clark [262]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Brink and Louw [429], COSMIN [426]
Clark [263]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Downs and Black [434], COSMIN [421]
Crowley [264]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	No	Not reported	Yes	Not reported	Not conducted	N/A
Cutolo [265]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	National Institute of Health (NIH) tool for observational cohort and cross-sectional studies (Appendix 11)
De Albuquerque [266]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS [422] and QAREL [423]
De Guio [267]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Partially (only start	Not reported	No	Not reported	Not conducted	N/A

					or end date given)					
De Langen [137]	Yes	Yes (tabulated)	MEDLINE plus other	No	No	Not reported	No	Not reported	Not conducted	N/A
De Paula Lima [268]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [426]
De Valk [269]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Not reported	National Institute of Health (NIH) tool for observational cohort and cross-sectional studies (Appendix 11)
Décary [270]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QAREL [423]
DeJong [271]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	No	Not reported	Not reported	QAREL [423]
Dekkers [272]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]
Deng [273]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Denteneer [274]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Not reported	Yes	Not reported	Independent and duplicate	COSMIN [134]
Denteneer [275]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [134]
D'hondt [276]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	No	Independent and duplicate	Independent and duplicate	COSMIN [421]

Dobson [277]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]
Dobson [278]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [134]
Ekpo [279]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Not reported	No	Independent and duplicate	Independent and duplicate	Viswanathan et al [435]
English [280]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Single reviewer	Independent and duplicate	Tool used in Pretorius and Keating [436]
Fan [281]	No	Yes (tabulated)	PUBMED only	Partially (terms in text only)	No	Not reported	No	Not reported	Not conducted	N/A
Ferreira [282]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QUADAS 2 [437]
Field [283]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Single reviewer	COSMIN [419] and McMaster rating (details not provided)
Fisher [284]	Yes	No	PUBMED plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Independent and duplicate	Cancer EDGE Task Force Rating Form [431] (Appendix A)
Flamand [285]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Single reviewer	Yes	Independent and duplicate	Independent and duplicate	Law and MacDermid [425]
Fonseca [286]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Single reviewer with checks by a second	Yes	Not reported	Independent and duplicate	Law and MacDermid [425]
Fotheringham [287]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Single reviewer with checks by a second	Not conducted	N/A

Gebruers [288]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Gengo e Silva [289]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Not reported	Yes	Not reported	Not conducted	N/A
Giraud [290]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Not reported	Yes	Not reported	Independent and duplicate	Not clear
Golden [291]	Yes	Yes (tabulated)	MEDLINE only	Yes (full strategy)	Partially (only start or end date given)	Not reported	No	Not reported	Not conducted	N/A
Gonzalez Lao [156]	Yes	Yes (tabulated)	Not reported	No	Yes	Not reported	No	Not reported	Independent and duplicate	BIVAC [153]
Gonzalez-Suarez [292]	Yes	Yes (tabulated)	PUBMED plus other	No	No	Not reported	No	Not reported	Not conducted	N/A
Guillaud [293]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Independent and duplicate	QAREL [423]
Hafsteinsdóttir [294]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Author's own
Hernaez [295]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Not reported	QUADAS [422] and STARD 2015 [438]
Himuro [296]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Single reviewer with checks by a second	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Hoogervorst [297]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Not reported	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]

Hulteen [298]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Tool used in Robertson et al [433]
Hunter [157]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Downs and Black [434]
Janaudis-Ferreira [299]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Single reviewer with checks by a second	Yes	Single reviewer	Independent and duplicate	COSMIN [134]
Jaspers [300]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	No	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [134]
Johnston [301]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]
Jonsson [302]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QAREL [423]
Jørgensen [303]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Kasehagen [304]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Single reviewer	Yes	Not reported	Independent and duplicate	Brink and Louw [429]
Kleijn [158]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [134]
Klingels [305]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Konieczka [306]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]

Kramer [138]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Not conducted	N/A
Kristensen [307]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Not reported	Yes	Not reported	Not conducted	N/A
Kroman [308]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]
Kwah [309]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Tool designed by Hebert et al [439]
Lagarde [310]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Single reviewer	Yes	Not reported	Independent and duplicate	The Dutch “Cochrane checklist for diagnostic accuracy studies” [440]
Lamers [311]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author’s own (presented in Appendix 1)
Lange [147]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	QAREL [423]
Lange [312]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	QAREL [423]
Larsen [313]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Single reviewer	Independent and duplicate	COSMIN [134]

Le Flao [314]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Not reported	Yes	Independent and duplicate	Independent and duplicate	STROBE [441]
Lemeunier [315]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	QAREL [423]
Lemeunier [316]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	QAREL [423]
Lennon [317]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Not reported	COSMIN [134]
Liang [318]	Yes	Yes (tabulated)	Not reported	No	No	Not reported	No	Not reported	Not conducted	N/A
Lima [319]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	No	Independent and duplicate	Yes	Not reported	Independent and duplicate	QUADAS [422]
Lisboa [320]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Tool used in Van Vlijmen et al [442]
Maaswinkel [321]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]
MacKay [322]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Single reviewer	QAREL [423]
Mahaffey [323]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	No	Not reported	Not reported	COSMIN [134]

Maricar [324]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Marshall [325]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer	Independent and duplicate	COSMIN [134]
May [326]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer for screening titles and abstracts, Independent and duplicate for full text assessment	Yes	Independent and duplicate	Independent and duplicate	Author's own (details presented in Table 1)
Mc Auliffe [327]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Single reviewer	Independent and duplicate	QAREL [423]
McCreesh [328]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QAREL [423]
McVerry [329]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Not reported	No	Not reported	Not conducted	N/A
Michiels [330]	Yes	Yes (tabulated)	PUBMED plus other	No	Yes	Not reported	Yes	Not reported	Independent and duplicate	CBO guidelines (details not provided)
Mieritz [331]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author's own
Mijnarends [332]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]

Milani [333]	Yes	Yes (tabulated)	PUBMED only	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Milne [334]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	No	Not reported	Not reported	COSMIN [134]
Mohan [335]	Yes	Yes (tabulated)	PUBMED plus other	No	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QAREL [423]
Mohseni Bandpei [336]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Moloney [337]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]
Moore [338]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS 2 [437]
Mulder-Brouwer [339]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	COSMIN [134]
Muntaner-Mas [340]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Not reported	Cochrane collaboration's tool [443] (domains analysed: detection bias, attrition bias, reporting bias, and other bias)
Nae [341]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	COSMIN [421]

Navarro [159]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	Brink and Louw [429]
Neto [342]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	No	Independent and duplicate	Not conducted	N/A
Nijholt [343]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Independent and duplicate	Tool used in Pretorius and Keating [436]
Oberwahrenbrock [344]	Yes	Yes (tabulated)	PUBMED only	Yes (full strategy)	Partially (only start or end date given)	Not reported	No	Not reported	Not conducted	N/A
O'Meara [345]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer with checks by a second	Yes	Single reviewer with checks by a second	Single reviewer with checks by a second	QUADAS [422]
Ornetti [346]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Not reported	Yes	Single reviewer	Not conducted	N/A
Ortega [347]	Yes	Yes (tabulated)	PUBMED plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Not conducted	N/A
Özay [348]	Yes	Yes (tabulated)	MEDLINE only	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Cochrane collaboration's tool [441]
Paech [349]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Not reported	No	Single reviewer	Not reported	Merlin et al [444] and QUADAS [422]
Pareira [350]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Single reviewer	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]

Parry [351]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	No	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]
Paul [352]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Not reported	Brughelli et al [445]
Peeling [353]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Independent and duplicate	Not reported	Author's own
Perdomo [354]	Yes	No	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Independent and duplicate	Independent and duplicate	Cancer EDGE Task Force Rating Form [431] (Appendix A)
Petitclerc [355]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	No	Independent and duplicate	Independent and duplicate	COSMIN [421]
Phillips [356]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Not reported	GRADE criteria (details not provided)
Pin [357]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	No	Single reviewer	Yes	Single reviewer	Single reviewer	COSMIN [421]
Pollock [358]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	No	Not reported	Not conducted	N/A
Ponce-Garcia [359]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	COSMIN [421]
Pons [360]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	No	Independent and duplicate	Independent and duplicate	Author's own
Pons [361]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author's own
Powden [139]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]

Proud [362]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [426]
Prowse [363]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Single reviewer for screening titles and abstracts, Independent and duplicate for full text assessment	Yes	Not reported	Independent and duplicate	Brink and Louw [429]
Rabelo [144]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Not reported	Brink and Louw [429]
Ratter [364]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Not reported	COSMIN [421]
Reavis [149]	Yes	Yes (tabulated)	MEDLINE only	Partially (terms in text only)	Partially (only start or end date given)	Not reported	No	Independent and duplicate	Not conducted	N/A
Reichmann [160]	Yes	Yes (tabulated)	EMBASE plus other	Partially (terms in text only)	Partially (only start or end date given)	Single reviewer	No	Single reviewer	Not conducted	N/A
Reis Durao [365]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS 2 [437]
Ringshausen [366]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Single reviewer with checks by a second	Not reported	Author's own

Roberts [367]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Partially (only start or end date given)	Not reported	No	Not reported	Not conducted	N/A
Robertson [368]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Independent and duplicate	Independent and duplicate	Author's own
Roeing [369]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Not reported	Yes	Not reported	Not conducted	N/A
Rondoni [370]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]
Rossini [371]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS 2 [437]
Rozema [151]	Yes	Yes (tabulated)	MEDLINE only	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Not conducted	N/A
Rubio-Ochoa [372]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QAREL [423]
Ruhe [373]	Yes	Yes (tabulated)	MEDLINE plus other	No	Yes	Independent and duplicate	No	Independent and duplicate	Single reviewer with checks by a second	Author's own
Rydwik [374]	Yes	Partially (narrative description in text)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Independent and duplicate	Independent and duplicate	COSMIN [426]
Saccomanno [375]	Yes	Partially (narrative description in text)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]

Saether [376]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]
Salamh [161]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer	Independent and duplicate	COSMIN [426]
Saltzherr [377]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Single reviewer	Independent and duplicate	QAREL [423]
Sam [378]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Unclear	Tool based on Bialocerkowski et al [432]
Scalco [379]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Not reported	STROBE [441]
Scheetz [380]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Single reviewer for screening titles and abstracts, Independent and duplicate for full text assessment	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]
Schrama [381]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author's own
Seagar [382]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]
Sepriano [383]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	National Institute of Health (NIH) tool for

										observational cohort and cross-sectional studies (Appendix 11)
Serai et al [152]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Not reported	QUADAS 2 [437]
Shiel [384]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	Brink and Louw [429]
Silva [385]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Tool used in van Bloemendaal et al [446] (based on COSMIN [134])
Simperingham [386]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Not conducted	N/A
Singh [387]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Not conducted	N/A
Sman [388]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS [422]
Smith [389]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Single reviewer with checks by a second	Critical Appraisal Skills Programme (CASP) tool (details not provided)
Smith [390]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Single reviewer with checks by a second	Critical Appraisal Skills Programme (CASP) tool (details not provided)
Sollis [391]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Independent and duplicate	Independent and duplicate	Tool based on the STARD Initiative [447] (details presented in Annex S1)

Southerest [392]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Independent and duplicate	QUADAS [422]
Stephenson [393]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Not reported	Not reported	Author's own
Stienen [394]	Yes	Partially (narrative description in text)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Not conducted	N/A
Stovall [395]	Yes	Yes (tabulated)	MEDLINE plus other	Partially (terms in text only)	No	Not reported	No	Not reported	Single reviewer with checks by a second	Author's own
Symonds [396]	Yes	No	MEDLINE only	Partially (terms in text only)	No	Not reported	No	Not reported	Not conducted	N/A
Tagmouti [28]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QAREL [423]
Talma [397]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Tarara [398]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer	Single reviewer	Tool based on COSMIN [134] and COSMIN [426]
Terwee [399]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]

Timmer [400]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Tong [401]	Yes	Yes (tabulated)	PUBMED only	Yes (full strategy)	Partially (only start or end date given)	Not reported	No	Not reported	Not conducted	N/A
Toohey [402]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Brink and Louw [429]
Traverso [403]	Yes	Yes (tabulated)	PUBMED only	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author's own
Valet [404]	Yes	No	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Not reported	Unclear	Author's own
van Bloemendaal [405]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	COSMIN [134]
van de Pol [406]	Yes	Yes (tabulated)	MEDLINE only	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author's own
van de Water [407]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	COSMIN [134]
van Kooij [408]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [134]
van Trijffel [409]	Yes	Yes (tabulated)	MEDLINE only	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Author's own

Wang [410]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Not reported	Yes	Not reported	Unclear	QUADAS 2 [437]
Weiner [29]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Yes	Single reviewer (10% duplicate)	Yes	Single reviewer (10% duplicate)	Not reported	Author's own
Welton [141]	Yes	Partially (narrative description in text)	MEDLINE plus other	Partially (terms in text only)	Yes	Not reported	Yes	Independent and duplicate	Independent and duplicate	Author's own
Wen [411]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Partially (only start or end date given)	Independent and duplicate	Yes	Not reported	Not reported	Tool used in Robertson et al [433]
Wetterslev [412]	Yes	Partially (narrative description in text)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Not conducted	N/A
Williams [413]	Yes	Partially (narrative description in text)	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer for screening titles and abstracts, Independent and duplicate for full text assessment	Yes	Single reviewer with checks by a second	Independent and duplicate	Author's own
Winser [414]	Yes	Partially (narrative description in text)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Wouters [415]	Yes	Partially (narrative	MEDLINE plus other	Yes (full strategy)	Yes	Single reviewer	Yes	Single reviewer	Not reported	COSMIN [421]

		description in text)				with checks by a second		with checks by a second		
Yang [416]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer	Independent and duplicate	COSMIN [421]
Yang [417]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Yes	Independent and duplicate	Yes	Single reviewer with checks by a second	Independent and duplicate	COSMIN [421]
Yoon [30]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS 2 [437]
Zanudin [418]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	COSMIN [421]
Zayat [419]	Yes	Yes (tabulated)	PUBMED plus other	Partially (terms in text only)	Yes	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	QUADAS 2 [437]
Zimmerman [420]	Yes	Yes (tabulated)	MEDLINE plus other	Yes (full strategy)	Partially (only start or end date given)	Independent and duplicate	Yes	Independent and duplicate	Independent and duplicate	Tool used in Alsufyani [448] "with minimal and appropriate adjustments" (details presented in Figure 1)

Appendix C: Search strategy and additional tables for Chapter 6

Appendix C1. Search strategy.

The search strategy for MEDLINE was:

1 Hand Strength/ (13203)

2 (hand adj2 grip$).ti,ab. (2764)

3 handgrip.ti,ab. (5351)

4 ((hand$ or grip$ or grasp$) and dynamom$).ti,ab. (2925)

5 (grip adj2 (strength or trial or pressure)).ti,ab. (9901)

6 muscle strength/ph and (hand$ or grip$ or grasp$).ti,ab. (1481)

7 muscle strength dynamometer/ and (hand$ or grip$ or grasp$).ti,ab. (621)

8 (((hand$ or muscle) adj2 (function or strength)) and (grip or grasp)).ti,ab. (3848)

9 ((jamar or takei or (grip adj it) or (eval adj solosystem) or biopac or harpenden) and dynamom$).ti,ab. (392)

10 or/1-9 (25231)

11 ((relative or absolute) adj reliability).ti,ab. (534)

12 exp Observer Variation/ (40491)

13 observer varia$.ti,ab. (3628)

14 ((inter or intra) adj (rater or tester or technician or examiner or assay or individual or participant)).ti,ab. (29571)

15 (interrater$ or intertester$ or intertechnician$ or interexaminer$ or interassay$ or interindividual$ or interparticipant$).ti,ab. (27049)

16 ((within or between) adj (subject or individual or person)).ti,ab. (28469)

17 exp "Reproducibility of Results"/ (375090)

18 (test and retest).ti,ab. (24015)

19 ((test or retest) and reliab$).ti,ab. (80228)

20 ((replicab$ or repeat$ or reproducibil$ or reliabil$ or valid$) and (measure$ or finding$ or result$ or test$)).ti,ab. (931492)

21 kappa$.ti,ab. (159950)

22 (intraclass adj correlation$).ti,ab. (20857)

23 ((individual or interval or rate) adj variability).ti,ab. (24733)

24 (uncertain$ adj5 measur$).ti,ab. (5759)

25 (coefficient$ adj3 varia$).ti,ab. (37432)

26 ((standard or technical) adj error$ adj2 measurement$).ti,ab. (2034)

27 (responsiv$ and (measure$ or finding$ or result$ or test$)).ti,ab. (139841)

28 ((minimal$ or clinical$ or small$ or least) adj2 (real or detectable or important or significant) adj2 (concentration or change$ or difference$)).ti,ab. (14869)

29 (limit$ adj2 (agreement$ or detection$)).ti,ab. (109291)

30 or/11-29 (1640920)

31 exp Animals/ (22216895)

32 exp Humans/ (17650455)

33 31 not 32 (4566440)

34 (10 and 30) not 33 (3607)

The search strategy for Embase was:

1 exp Hand grip/ (9542)

2 (hand adj2 grip$).ti,ab. (4532)

3 handgrip.ti,ab. (7646)

4 ((hand$ or grip$ or grasp$) and dynamom$).ti,ab. (4980)

5 (grip adj2 (strength or trial or pressure)).ti,ab. (14292)

6 exp dynamometer/ and (hand$ or grip$ or grasp$).ti,ab. (2747)

7 (((hand$ or muscle) adj2 (function or strength)) and (grip or grasp)).ti,ab. (6187)

8 ((jamar or takei or (grip adj it) or (eval adj solosystem) or biopac or harpenden) and dynamom$).ti,ab. (749)

9 ((relative or absolute) adj reliability).ti,ab. (632)

10 exp Observer Variation/ (19384)

11 observer varia$.ti,ab. (6112)

12 ((inter or intra) adj (rater or tester or technician or examiner or assay or individual or participant)).ti,ab. (42846)

13 (interrater$ or intertester$ or intertechnician$ or interexaminer$ or interassay$ or interindividual$ or interparticipant$).ti,ab. (33230)

14 ((within or between) adj (subject or individual or person)).ti,ab. (35813)

15 exp "reproducibility"/ (201596)

16 (test and retest).ti,ab. (29872)

17 ((test or retest) and reliab$).ti,ab. (108772)

18 ((replicab$ or repeat$ or reproducibil$ or reliabil$ or valid$) and (measure$ or finding$ or result$ or test$)).ti,ab. (1336782)

19 kappa$.ti,ab. (185876)

20 (intraclass adj correlation$).ti,ab. (25965)

21 ((individual or interval or rate) adj variability).ti,ab. (35192)

22 (uncertain$ adj5 measur$).ti,ab. (6609)

23 (coefficient$ adj3 varia$).ti,ab. (47805)

24 ((standard or technical) adj error$ adj2 measurement$).ti,ab. (2469)

25 (responsiv$ and (measure$ or finding$ or result$ or test$)).ti,ab. (180070)

26 ((minimal$ or clinical$ or small$ or least) adj2 (real or detectable or important or significant) adj2 (concentration or change$ or difference$)).ti,ab. (22128)

27 (limit$ adj2 (agreement$ or detection$)).ti,ab. (131215)

28 exp animal/ (23895646)

29 exp human/ (19485201)

30 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 (29825)

31 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 (2050476)

32 28 not 29 (4410445)

33 30 and 31 (4630)

34 33 not 32 (4413)

Appendix C2. Data extraction form.

Items extracted from the identified studies	
A: Study characteristics	
A1	Study design
A2	Sample size
A3	Data collection
A4	Type of reliability
A5	Number of testing sessions
A6	Time interval between testing sessions
A7	Number of measurements made within testing sessions
A8	Time interval between measurements made within testing sessions
B: Observer characteristics	
B1	Number of participating observers
B2	Experience of observers
C: Participant characteristics	
C1	Medical condition
C2	Age of participants
C3	Gender of participants
C4	Inclusion criteria
D: Measurement conditions	
D1	Device calibrated
D2	Preparatory instructions provided to participants
D3	Vocal encouragement provided
E: Measurement protocol	
E1	Summary measure per session
E2	Tested hand(s)
E3	Dynamometer used
E4	Handle position
E5	Posture
E6	Positioning of shoulder/elbow/wrist
F: Statistical estimates	
F1	Statistical estimates reported for reliability
F2	Statistical estimates reported for measurement error

Appendix C3. Standards on design requirements and statistical methods for studies on reliability or measurement error. After Mokkink et al [37].

	Design requirements	very good	adequate	doubtful	inadequate
1	Were patients stable in the time between the administration of the repeated measurements on the construct to be measured?	Evidence provided that patients were stable	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
2	Was the time interval between the measurements performed within testing sessions appropriate?	Time interval higher than or equal to 30 seconds	Time interval between 15 and 30 seconds	Time interval less than 15 seconds	No rest given or not stated
3	Was the time interval between testing sessions appropriate?	Evidence provided that time interval was appropriate	Assumable that time interval was appropriate	Doubtful whether time interval was appropriate or time interval was not stated	Time interval not appropriate
4	Did the professional(s) administer the measurement without knowledge of scores of other repeated measurement(s) in the same patients?	Measurements administered without knowledge of other scores (evidence provided)	Assumable that measurements were administered without knowledge of other scores	Unclear if measurements were administered without knowledge of other scores	Measurements administered with knowledge of other scores

5	Did the professional(s) assign scores or determine values without knowledge of the scores of other repeated measurement(s) in the same patients?	Scores assigned or values determined without knowledge of other scores (evidence provided)	Assumable that scores were assigned or values were determined without knowledge of other scores	Unclear if scores were assigned or values were determined without knowledge of other scores	Scores were not assigned or values were not determined without knowledge of other scores
6	Were the statistical parameters used for reliability appropriate?	ICC or G Coefficient calculated; the model or formula was described	ICC or G Coefficient calculated but model or formula was not described OR Pearson or Spearman correlation coefficient calculated		Statistical parameters used not optimally matching the research question
7	Were the statistical parameters used for measurement error appropriate?	SEM, SDC, LoA or CV calculated; the model or formula for the SEM/SDC is described	SEM, SDC, LoA or CV calculated, but the model or formula is not described		Statistical parameters used not optimally matching the research question

[1]Based on Mathiowetz et al [198], [2]Pearson and Spearman correlation coefficients were considered adequate, as no systematic differences were expected due to the objective nature of the test

Appendix C4. Data extraction - Characteristics of the identified primary studies.

Author	Year	Population	Sample size	Data collection	Type of variability examined	Number of testing sessions	Time interval between testing sessions	Number of within-session measurements	Time interval between within-session measurements
Abizanda [449]	2012	Elderly	281	Prospective	Test-retest	One	N/A (one session)	Three	15 seconds
Aguiar [450]	2016	Patients in the subacute phase of stroke	32	Prospective	Inter-observer & test-retest	Two	1-2 weeks	Three	15-20 seconds
Alencar [183]	2012	Elderly patients with dementia	72	Prospective	Test-retest	Two	1 week	Three	1 minute
Alfonso-Rosa [129]	2014	Older adults with Type 2 Diabetes	16	Prospective	Test-retest	Two	1 week	Three	1 minute
Allen [451]	2011	Healthy participants	49	Prospective; convenience sample	Test-retest	Two	1 week	Three	15 seconds
Anumula [452]	2014	University students	27	Prospective	Test-retest	Two	1 week	Three	15-60 seconds
Baldwin [453]	2013	Critically ill & healthy participants	17 critically ill, 12 healthy	Prospective, consecutive	Inter-observer & test-retest	Two	2 days	Three	1 minute
Barden [191]	2012	Adults with upper motor neuron syndrome	36 adults with acquired brain injury, and 27 healthy individuals	Prospective; consecutive	Test-retest	Two	5 weeks	Ten	Not clear ("consecutive")
Bertrand [454]	2015	Subjects who had a stroke	34	Prospective; convenience sample	Test-retest	Two	Same week	Three	Not clear ("consecutive")
Blankevoort [455]	2013	Older people with dementia	57	Prospective; convenience sample	Test-retest	Two	1 week	Three	Not given

Bodilsen [456]	2015	Acutely admitted older medical patients	52	Prospective; random sample	Inter-observer	Two	2-3 hours	Three	Not given
Bohannon [457]	2005	Healthy community-dwelling elders	21	Prospective	Test-retest	Two	12 weeks	One	N/A
Bohannon [458]	2006	Healthy volunteers	30	Prospective	Test-retest	Two	1 day	Two	30 seconds
Bohannon [459]	2011	Healthy participants	28	Prospective	Test-retest	Two	4-10 days	Two	Not given
Boissy [460]	1999	Chronic stroke and healthy subjects	15 stroke and 10 healthy patients	Prospective	Test-retest	Two	1 week	Three	2 minutes
Brogardh [461]	2015	Patients with late effects of polio in their upper limbs	28	Prospective	Test-retest	Two	14 days	Two	90 seconds
Brown [462]	2000	Hand-injured patients	30	Prospective	Inter-observer & test-retest	Two	"The time between therapists was standardized to the time it took the patient to switch testing stations"	Three	5 seconds
Buehring [463]	2014	Older community dwelling individuals	97	Prospective	Test-retest	Two	3 months	Three	10-20 seconds
Burnstein [464]	2011	Healthy participants	238	Retrospective	Test-retest	Two	6 months	Two	Not given
Carbonell-Baeza [465]	2015	Females with fibromyalgia	100	Prospective; convenience sample	Test-retest	Two	7 days	Three	1 minute
Chen [466]	2009	Patients with history of sustained stroke	62	Prospective; convenience sample	Test-retest	Two	3-7 days	Three	Not given

Clifford [113]	2013	Patients with UL burns	89	Prospective	Test-retest	Four	Sessions took place at 1, 3, 6, 12 months	Three	Not given
Coldman [182]	2006	Subjects with post carpal tunnel decompression, post flexor tendon repair, and asymptomatic subjects.	66	Prospective	Test-retest	Two	15 minutes	Three	15 seconds
Dunn [186]	1978	Mentally retarded adult males	24	Prospective; random sample	Test-retest	Four	1 day (for 4 consecutive days)	Two	Not given
Ekstrand [467]	2015	Subjects with paresis in upper extremity 6 months post-stroke	45	Prospective	Test-retest	Two	1 week	Three	60 seconds
Essendrop [468]	2001	Healthy subjects	19	Prospective	Test-retest	Two	1 week	Three	30 seconds
Fox [469]	2014	Older adults with dementia	12	Prospective	Test-retest	Two	7 days	Not clear	N/A
Gatt [470]	2018	Subjects with hand & wrist injuries	160	Retrospective	Test-retest	One	N/A (one session)	Three	None
Gerhardsson [471]	2014	Subjects with hand-arm vibration syndrome (HAVS)	47	Retrospective	Test-retest	Two	2 weeks	Three	Not given
Gerodimos [472]	2012	Basketball players	30	Prospective	Test-retest	Two	1 day	Three	60 seconds
Gittings [473]	2016	Patients with unhealed minor burn wounds	30	Prospective	Test-retest	One	N/A (one session)	Three	Not given
Gittings [474]	2018	Burn-injured patients	38	Prospective	Test-retest	One	N/A (one session)	Three	Not given

Guerra [196]	2017	Group 1: Sample of free-living subjects among student and staff from the Uni of Porto, and relatives / Group 2: Sample of inpatients recruited from a university hospital in Porto	164	Prospective; consecutive	Test-retest	One	N/A (one session)	Two	1 minute
Haidar [475]	2004	Hospital worker volunteers	100	Prospective	Test-retest	Two	2 weeks	Three	1 minute
Hamilton [188]	1992	Healthy female college-age subjects	29	Prospective	Test-retest	Two	1 week	Three	>=4 minutes
Hamilton [181]	1994	Healthy subjects	33	Prospective	Test-retest	Two	12 days on average [range: 1-22 days]	Three	Not given
Haward [476]	2002	Healthy subjects	18	Prospective	Test-retest	Two	1 week	Three	10 seconds
Hilgenkamp [477]	2012	Older adults with intellectual disability	36	Retrospective	Test-retest	Two	2 weeks	One	N/A
Huang [189]	2011	Patients with stroke	56	Retrospective	Test-retest	Two	3-7 days	Three	Not given
Irwin [478]	2010	University students & older community dwelling individuals	28	Prospective	Test-retest	Two	1-2 weeks	Three	1 minute
Jenkins [479]	2017	Sarcopenic subjects	257	Retrospective	Test-retest	Two	Not clear	Three	Not given

Kennedy [480]	2010	Patients with rheumatoid arthritis	25	Prospective	Test-retest	Two	Not given. There was a 15-min break between the four different testing conditions, but no mention of the actual test-retest interval	Three	1 minute
Khamwong [481]	2010	Healthy men	25	Prospective	Test-retest	Two	1 day	Three	1 minute
MacDermid [482]	1994	Patients with cumulative trauma disorders	38	Prospective	Inter-observer	Two	1 hour	Three	15 seconds
Maher [483]	2018	Healthy adults	103	Prospective	Inter-observer & test-retest	Two	1 min	Three	None
Mawdsley [484]	2001	Elderly females	23	Prospective	Inter-observer	Two	15 min	Three	"The amount of time it took the tester to read and record the data after each trial served as the rest period between trials."
Medina-Mirapeix [485]	2016	Patients with COPD	30	Prospective	Inter-observer	Two	7-14 days	Three	N/A
Niebuhr [486]	1994	Healthy students	33	Prospective	Test-retest	Two	12-30 days	Three	30 seconds
Nitschke [193]	1999	Healthy & disabled women	32 healthy and 10 with non-specific regional pain	Prospective	Test-retest	Two	4-7 days	Three	20 seconds
Paltamaa [487]	2005	Multiple sclerosis	28 in total (19 for test-retest, 9 inter-rater)	Prospective	Inter-observer & test-retest	Two	1 week	Three	Not given

Peolsson [488]	2001	Healthy volunteers & patients with cervical radiculopathy	32 healthy and 13 with cervical radiculopathy	Prospective; convenience sample	Inter-observer & test-retest	Three	at least 1 day	Three	Not given
Plant [489]	2016	Healthy adults	25	Prospective	Inter-observer & test-retest	Two	minimum 1 week	Three	15 seconds
Puthoff [490]	2013	Cardiac rehabilitation	49	Prospective	Test-retest	Two	15 minutes	Three	Not given
Reddon [187]	1985	Self-reported dextral subjects	12	Prospective	Test-retest	Ten	1 week	Ten	30 seconds
Reijnierse [195]	2017	Low muscle strength (Dynapenia)	939	Retrospective	Test-retest	One	N/A (one session)	Three	Not given
Reuter [184]	2011	Healthy adults	21	Prospective	Test-retest	Three	3 sessions separated by at least 24 hours	Three	1 minute
Savva [491]	2013	Healthy students	19	Prospective	Test-retest	Two	7 days	Three	15 seconds
Savva [33]	2014	Patients with cervical radiculopathy	19	Prospective	Test-retest	Two	7 days	Three	15 seconds
Savva [128]	2018	Patients with shoulder impingement syndrome	19	Prospective	Test-retest	Two	7 days	Three	15 seconds
Schaubert [185]	2005	Community-dwelling elderly persons	10	Prospective; convenience sample	Test-retest	Three	every 6 weeks (baseline, then at 6, 12 weeks)	Not clear	1 minute
Schreuders [492]	2003	Subjects with hand injuries	33	Prospective; consecutive sample	Inter-observer & test-retest	Two	2-3 minutes	Three	Not given
Segura-Orti [493]	2011	Subjects undergoing hemodialysis	12	Prospective	Test-retest	Two	1 to 2 weeks	Three	15 seconds
Shechtman [179]	2003	Healthy subjects	180	Prospective	Test-retest	Two	15 minutes	Three	30 seconds

Shechtman [494]	2005	Healthy adults	100	Prospective	Test-retest	Two	10 minutes	Three	30 seconds
Silva [495]	2019	Ambulatory older adults	100	Prospective	Inter-observer	Two	2-7 days	Three	Not given
Smidt [496]	2002	Patients with Lateral Epicondylitis	50	Retrospective	Inter-observer	Two	Not given	Three	20 seconds
Solari [497]	2008	Patients with Charcot-Marie-Tooth disease	Phase I: 40, Phase II: 26	Prospective	Inter-observer & test-retest	Two	1 week	Three	10 seconds
Spijkerman [498]	1991	Healthy subjects & subjects with impaired hand function	24	Prospective	Test-retest	Two	1 week	Three	1 minute
Stephens [499]	1996	Subjects without upper extremity abnormality and subjects following open-palm carpal tunnel release (CTR)	78	Prospective	Inter-observer (between-session) & test-retest (within-session)	Two	1 day	Three	30 seconds
Stockton [500]	2011	Women with systemic lupus erythematosus	12	Prospective	Test-retest	Two	7-10 days	Three	30 seconds
Svensson [501]	2006	Adults with Charcot-Marie-Tooth	20	Prospective	Test-retest	Two	1 week	Three	1 minute
Tager [502]	1998	Older population	199	Prospective; consecutive	Test-retest	Two	48 hours	Not clear	several minutes
Tan [503]	2001	Healthy participants	39	Prospective	Test-retest	Two	1 day	Three	25 seconds minimum
Trippolini [504]	2013	Patients with Whiplash-associated disorders	32	Prospective; convenience sample	Test-retest	Two	1 week	Three	Not given

Trutschnigg [505]	2008	Advance cancer patients	74	Prospective	Test-retest	Two	Not clear (says same day, and 1 minute between the 2 different devices)	Three	consecutive
Tsang [506]	2005	Healthy volunteers	548	Prospective; convenience sample	Test-retest	Two	3 days	Three	15-20 seconds
Tveter [507]	2014	Subjects with musculoskeletal conditions	81	Prospective	Test-retest	Two	1 week	Two	Not given
Vermeulen [508]	2015	Older adults	88	Prospective	Test-retest	One	N/A (one session)	Three	Not clear (“consecutive”)
Villafane [509]	2015	Subjects with thumb carpometacarpal osteoarthritis	78	Prospective; convenience sample	Test-retest	Two	1 week	Three	1 minute
Villafane [510]	2016	Subjects with Parkinson disease (PD) and healthy subjects	30 (15 with PD and 15 healthy)	Prospective; convenience sample	Test-retest	Two	1 week	Three	1 minute

Appendix C5. Data extraction (continued) - Characteristics of study individuals and participating observers.

Author	Year	Characteristics of study individuals			Characteristics of participating observers	
		Age	Gender (N men, %)	Underlying medical conditions	Number of observers	Where the observers experienced?
Abizanda [449]	2012	Mean=74.3 (SD=4.9)	103 (37%)	Arterial hypertension, Dyslipidaemia, Osteoarthritis, Depression, Diabetes	One	Not given
Aguiar [450]	2016	Mean=63 (SD=12)	18 (56%)	Stroke	Two	Yes (1 year plus training prior to assessments)
Alencar [183]	2012	Mean=83.9 (SD=5.8)	12 (16%)	Dementia	One	Not given
Alfonso-Rosa [129]	2014	Mean=73.6 (SD=8.1)	10 (56%)	Diabetes	One	Yes
Allen [451]	2011	Aged between 18 and 25yrs	7 (14%)	None	One	Yes
Anumula [452]	2014	Not given	Not given	None	One	Not given
Baldwin [453]	2013	Critically ill group: median=78 (Q1=46, Q3=82) / healthy group: median=55 (Q1=26, Q3=59)	Critically ill: 10 (59%) / healthy: 6 (50%)	Intra-abdominal sepsis/pancreatitis/multiple organ failure, Pneumonia, Respiratory failure, Cardiac failure/infarction, Cardiac surgery, Complicated drug reaction, Trauma	Two	Not given
Barden [191]	2012	ABI group: mean=50 (SD=15)/ Control: mean=40 (SD=12)	ABI group: 20 (56%), Control group: 14 (52%)	Stroke, traumatic brain injury	One	Not given
Bertrand [454]	2015	Mean=56.9 (SD=13.7)	18 (53%)	Stroke	One	Not given

Blankevoort [455]	2013	Mean=82.47 (SD=5.31)	17 (30%)	Older dementia patients	One	No
Bodilsen [456]	2015	Mean=78 (SD=8.3)	14 (27%)	Acutely admitted older patients	Three	Mixture (2 experienced and 1 newly qualified physiotherapist)
Bohannon [457]	2005	Mean=75 (SD=5.9)	4 (19%)	None	One	Not given
Bohannon [458]	2006	Mean=38 (SD=15.6)	14 (47%)	None	One	Not given
Bohannon [459]	2011	Mean=45.7 (SD=23.5)	14 (50%)	Not stated	One	Not given
Boissy [460]	1999	Stroke subjects: mean=47 (SD=14) / Controls: mean=44 (SD=11)	Stroke subjects: 10 males (67%), Controls: 5 (50%)	Stroke	One	Yes
Brogardh [461]	2015	Mean=68 (SD=11)	16 (57%)	Polio	One	Not given
Brown [462]	2000	Mean=43 (SD=12.3)	22 (73%)	Hand injuries	Two	Yes
Buehring [463]	2014	Mean=80.7yrs (range 70-95)	48 (49%)	None	One	Not given
Burnstein [464]	2011	Mean=28.7 (SD=6.4)	162 (68%)	Not reported	One	Yes
Carbonell-Baeza [465]	2015	Mean=50.6 (SD=8.6)	0 (0%)	Fibromyalgia	One	Not given
Chen [466]	2009	Mean=61.0 (SD=9.9)	45 (73%)	Stroke	One	Not given
Clifford [113]	2013	Mean=35.5 (SD=14.5)	75 (84.3%)	UL burn	One	Yes
Coldman [182]	2006	Asymptomatic: mean=40.4yrs (range 23-72) / Carpal tunnel decompressions:	Asymptomatic: 11 (50%) / Carpal tunnel decompressions: 6 (27%) / Flexor	Hand impairments	One	Yes (5 years)

		mean=60.5yrs (range 38-93) / Flexor tendon repairs: mean=39.9yrs (range 20-72)	tendon repairs: 16 (73%)			
Dunn [186]	1978	Mean=24.21yrs (range 18-59)	24 (100%)	Mental retardation	One	Not given
Ekstrand [467]	2015	Mean=65 (SD=7)	37 (82%)	Stroke	One	Yes
Essendrop [468]	2001	Mean=35 (SD=6.9)	6 (32%)	None	One	Not given
Fox [469]	2014	Mean=83.25 (SD=9.94)	1 (8%)	Dementia	One	Yes
Gatt [470]	2018	Mean=23.7 (SD=3.9)	129 (81%)	Hand & wrist injuries	One	Not given
Gerhardsson [471]	2014	HAVS group: mean=50.4 (SD=12.4) / Reference group: mean=37.6 (SD=15.9)	HAVS group: 36 (77%) / Reference group: Not given	Hand-arm vibration syndrome (HAVS) or none	One	Yes
Gerodimos [472]	2012	Mean=26.06 (SD=5.57)	30 (100%)	None	One	Not given
Gittings [473]	2016	Median [IQR]: 28.5 [20.0]	25 (83.3%)	Burn injury	One	Not given
Gittings [474]	2018	Median=30 (Q1=23, Q3=39)	33 (87%)	Burn injury	One	Not given
Guerra [196]	2017	Group 1: mean=31 (SD=12) / Group 2: mean=76 (SD=8)	60 (37%)	Group 1: None / Group 2: Recruited from cardiology, endocrinology, gastroenterology, internal medicine, orthopaedics, and otolaryngology wards.	One	Not given
Haidar [475]	2004	Mean=35.5 (range 21-63)	50 (50%)	None	One	Not given

Hamilton [188]	1992	Mean=23.8 (SD=4.9)	0 (0%)	None	One	Not given
Hamilton [181]	1994	Mean=36.5 (range 20-55)	16 (48%)	None	One	Not given
Haward [476]	2002	Aged between 18 and 25yrs	18 (100%)	None	One	Not given
Hilgenkamp [477]	2012	Mean=66.8 (SD=9.9)	12 (33%)	Intellectual disabilities	One	Not given
Huang [189]	2011	Mean=61.9 (SD=9.3)	41 (73%)	Stroke	One	Not given
Irwin [478]	2010	Younger group: mean=23.4 (SD=3.8) / Older group: mean=75.6 (SD=7.0)	8 (29%)	None	One	Not given
Jenkins [479]	2017	Mean=76.4 (SD=6.5)	98 (38%)	Sarcopenia	One	Not given
Kennedy [480]	2010	Mean=62.5 (SD=10.6)	5 (20%)	None	One	Not given
Khamwong [481]	2010	Mean=20.6 (SD=1.3)	25 (100%)	None	One	Not given
MacDermid [482]	1994	Mean=41 (SD=14)	Not given	Cumulative trauma disorders	Two	Yes
Maher [483]	2018	Aged between 18 and 65yrs	30 (29%)	None	Two	No
Mawdsley [484]	2001	Mean=76.1 (SD=7.2)	0 (0%)	None	One	No
Medina-Mirapeix [485]	2016	Mean=67 (SD=6.49)	30 (100%)	Stable COPD	Two	Not given
Niebuhr [486]	1994	Mean=24.4 (SD=5.4)	5 (15%)	None	One	Not given
Nitschke [193]	1999	Healthy group: mean=32.3 (SD=7.3) /	0 (0%)	Non-specific regional pain	One	Not given

		NSRP group: mean=42.6 (SD=11.8)				
Paltamaa [487]	2005	Test-retest group: mean=42.7 (SD=9.2) / Inter-rater group: mean=48.9 (SD=8.8)	Test-retest group: 10 (53%) / Inter-rater group: 3 (33%)	Multiple sclerosis	Two	Yes
Peolsson [488]	2001	Healthy group: mean=29 (SD=10) / Diseased group: mean=50 (SD=12)	Healthy group: 8 (22%) / Diseased group: 7 (54%)	None or cervical radiculopathy	Three	Not given
Plant [489]	2016	Mean= 40yrs	10 (40%)	None	Two	No
Puthoff [490]	2013	Mean=68.7 (SD8.8)	29 (59%)	Cardiac events	One	Not given
Reddon [187]	1985	Men: 21 to 36yrs, women: 20 to 31yrs	6 (50%)	None	One	No
Reijnierse [195]	2017	63.3yrs (mean of all cohorts)	382 (41%)	Dynapenia	One	Not given
Reuter [184]	2011	Mean=29.8 (SD=6.9)	12 (52%)	None	One	Not given
Savva [491]	2013	Males: range=21-26, SD=2.5yrs / females: range=21-23, SD=1yr	10 (53%)	None	One	Yes
Savva [33]	2014	Mean=50.5 (SD=12.0)	14 (74%)	Cervical radiculopathy	One	Yes
Savva [128]	2018	Mean=33.2 (SD=12.9)	9 (47%)	Shoulder impingement syndrome	One	Not given
Schaubert [185]	2005	Mean=75.5 (SD=5.8)	(20%)	None	One	Not given

Schreuders [492]	2003	Mean=36 (SD=13.7)	20 (61%)	Hand injuries	Two	Mixture (1+1)
Segura-Orti [493]	2011	Not given	Not given	End-stage renal disease	One	Not given
Shechtman [179]	2003	Aged between 18 and 49yrs	80 (44%)	None	One	Not given
Shechtman [494]	2005	Mean=23.5 (SD=3.5)	50 (50%)	None	One	Not given
Silva [495]	2019	Mean=82.3 (SD=8.1)	38 (38%)	Depression & disability	Two	Both had at least 1 year of experience
Smidt [496]	2002	Mean=47 (SD=11)	30 (60%)	Lateral Epicondylitis	Two	Yes
Solari [497]	2008	Phase I: mean=42.4 (SD=12.6) / Phase II: mean=43.9 (SD=13.5)	Phase I: 19 (48%) / Phase II: 13 (50%)	Charcot-Marie-Tooth disease	Two	Not given
Spijkerman [498]	1991	Mean=29.5 (SD=10.2)	Not given	None or impaired hand function	One	Not given
Stephens [499]	1996	Surgical: Mean=51.23 (SD=13.93) / Non-surgical: mean=32.40 (SD=10.51)	Not given	Carpal tunnel release	Two	No
Stockton [500]	2011	Mean=39.8 (SD=10.0)	0 (0%)	Systemic lupus erythematosus	One	Yes
Svensson [501]	2006	Mean=51.2 (SD=13.9)	9 (45%)	Charcot-Marie-Tooth disease	One	Not given
Tager [502]	1998	Aged between 55 and 69yrs	100 (50%)	None	One	Not given
Tan [503]	2001	Mean=34.3 (SD=8.2)	26 (67%)	None	One	Not given
Trippolini [504]	2013	Mean=39.6 (SD=12.3)	21 (66%)	Whiplash-associated disorders	One	Yes

Trutschnigg [505]	2008	Mean=61.5 (SD=13.1)	48 (65%)	Cancer	One	Not given
Tsang [506]	2005	Mean=37.8 (SD=10.9)	226 (41%)	None	One	Not given
Tveter [507]	2014	Mean=57.6 (SD=14.2)	23 (28%)	Musculoskeletal conditions	One	Yes
Vermeulen [508]	2015	Mean=75.0 (SD=6.8)	33 (38%)	None	One	Not given
Villafane [509]	2015	Mean=83 (SD=5)	Not clear. Different numbers of females reported in abstracts and results, and there is no Table 1.	Osteoarthritis	One	Yes
Villafane [510]	2016	PD group: mean=69.5 (SD=8.6) / Healthy: mean=67.5 (SD=10.2)	PD group: 7 (47%) / Healthy: 6 (40%)	Parkinson's disease & none	One	Yes

Appendix C6. Data extraction (continued) - Measurement conditions and protocol used.

Author	Year	Measurement conditions			Measurement protocol					
		Device calibrated	Preparatory instructions provided to individuals	Vocal encouragement provided to individuals	Tested hand(s)	Session summary measure(s) used	Device(s) used	Handle position(s) used	Position of shoulder/forearm/ elbow/wrist	Posture of individuals
Abizanda [449]	2012	Not stated	Not stated	Not stated	Dominant	N/A (single session)	JAMAR	Not stated	Not stated	Sitting
Aguiar [450]	2016	Yes	Yes	Yes	Affected & contralateral	First & mean of 2 & mean of 3	SAEHAN	Not stated	As per ASHT[1] [123] recommendations	Sitting
Alencar [183]	2012	Not stated	Yes	Yes	Dominant	Mean of 3	JAMAR	Adjusted (second for women, third for men)	As per ASHT[1] [123] recommendations	Sitting
Alfonso-Rosa [129]	2014	Not stated	Yes	Yes	Dominant & non-dominant & bimanual	Mean of 3	Takei	Adjusted to the individual's hand size	Arm in complete extension. No other details provided.	Standing
Allen [451]	2011	Yes	Yes	Yes	Right & left	Mean of 3	Biometrics E-LINK EP9 electronic dynamometer, JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Anumula [452]	2014	Not clear	Yes	Not stated	Not clear	Highest of 3	Takei	Not stated	The participants were told to raise their arm above the head sideways without putting any pressure on the dynamometer, then bring the arm down with elbows fully extended in sideways position	Standing

									and simultaneously exert maximum force	
Baldwin [453]	2013	Yes	Yes	Yes	Right & left	Mean of 3	JAMAR	Second	Not stated	Sitting
Barden [191]	2012	Yes	Yes	Yes	Dominant & non-dominant	Mean of 10	JAMAR	Adjusted to the individual's hand size	As per ASHT[1] [123] recommendations, with a minor modification: the elbow and forearm were supported on the armrest of the chair or wheelchair to ensure appropriate shoulder support for participants with UMN lesions.	Sitting
Bertrand [454]	2015	Yes	Yes	Not given	Affected & contralateral	Mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Blankevoort [455]	2013	Not clear	Yes	Not clear	Dominant	Highest of 3	JAMAR	Not stated	The individuals had their arm extended and the palm of their hand was facing their leg. No other details provided.	Standing
Bodilsen [456]	2015	Not clear	Yes	Not clear	Dominant	Highest of 3	Saehan, Digi-II	Not stated	For patients able to leave the bed sat in a chair, elbows were flexed at 90 degree, the lower arm was placed on the armrest with the wrist in neutral position.	Sitting

Bohannon [457]	2005	Yes	Not stated	Not stated	Right & left	Single measurement	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Bohannon [458]	2006	Yes	Not clear	Yes	Right & left	First & mean of 2 & highest of 2	Micro FET 4	Not stated	As per ASHT[1] [123] recommendations	Sitting
Bohannon [459]	2011	Yes	Not clear	Not clear	Right & left	First & mean of 2 & highest of 2	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Boissy [460]	1999	Not clear	Not clear	Not clear	Right & left / Affected & contralateral	Highest of 3	Modified prehension dynamometer	Not stated	The shoulder was placed at approximately 30 degrees of adduction and 0 degrees of flexion. The elbow was flexed at 90 degrees with the wrist in neutral position.	Sitting
Brogardh [461]	2015	Yes	Yes	Yes	More affected & less affected	Highest of 2	GRIP-it	Not stated	The participants were seated with relaxed shoulders, and the tested forearm was supported with a soft pad. When the participants grasped the GRIP-it, the wrist was held in a position between 0-15 degrees dorsiflexion.	Sitting
Brown [462]	2000	Not clear	Yes	Yes	Affected	Mean of 3	JAMAR	Not stated	As per ASHT[1] [123] recommendations	Sitting
Buehring [463]	2014	Not clear	Not clear	Not clear	Non-dominant	Highest of 3	JAMAR	Not stated	Not stated	Standing

Burnstein [464]	2011	Not stated	Yes	Yes	Right & left	Highest of 2	Baseline	Adjusted to the individual's hand size	Shoulder adducted and in neutral rotation, elbow was flexed to 90°, and lower arm and wrist in neutral position.	Standing
Carbonell-Baeza [465]	2015	Not stated	Not clear	Not clear	Both hands (best of two measurements chosen for each hand, average then calculated)	Mean of 2 hands	Takei	Not stated	Arm fully extended, forming a 30 degree angle in relation to the trunk. No other details provided.	Not stated
Chen [466]	2009	Not stated	Yes	Yes	More affected & less affected	Mean of 3	Eval Solosystem digital dynamometer	Second	The shoulder was adducted, the elbow flexed at 90 degrees, and the forearm and wrist in neutral position.	Sitting
Clifford [113]	2013	Yes	Yes	Not clear	Right & left	Mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Coldman [182]	2006	Yes	Yes	Yes	Dominant & affected	First & mean of 3 & highest of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Dunn [186]	1978	Not given	Yes	No	Dominant & non-dominant	Mean of 2	Not given	Not stated	Not stated	Not stated
Ekstrand [467]	2015	Yes	Yes	Yes	More affected & less affected	Highest of 3	Grippit	Not stated	Forearm in neutral position, shoulder in 30 degrees, elbow in 90 degrees, wrist in 0 to 15 degrees dorsiflexion	Sitting
Essendrop [468]	2001	Yes	Yes	Yes	Right	Highest of 3	JAMAR	Not stated	The upper arm was kept vertical with a 90 degree flexion in the elbow. The palm	Sitting

									of the hand was held vertical. No other details provided.	
Fox [469]	2014	Not clear	Yes	Yes	Right & left	Not clear	JAMAR	Not stated	Participants were seated with their elbows at their sides and at 90 degrees. No other details provided.	Sitting
Gatt [470]	2018	Not clear	No	No	Right & left	N/A (single session)	Takei GRIP-D	Not stated	Arm by participant's side with full elbow extension. No other details provided.	Standing
Gerhardsson [471]	2014	Yes	No	No	Right & left	Mean of 3	Baseline	Second	Not stated	Not stated
Gerodimos [472]	2012	Not given	Yes	Not given	Dominant & non-dominant	Highest of 3	JAMAR	Not stated	The shoulder of tested arm was adducted, the elbow flexed at 90 degrees, whereas the forearm and wrist were set in neutral position.	Sitting
Gittings [473]	2016	Not given	Yes	Not given	Not clear	N/A (single session)	JAMAR	Not stated	The shoulder was adducted and neutrally rotated, elbow at 90 degrees flexion, forearm in neutral position and wrist between 0 and 30 degrees flexion and between 0 and 15 degrees ulnar deviation.	Sitting
Gittings [474]	2018	Not given	Yes	Yes	Right & left	N/A (single session)	JAMAR	Not stated	Shoulder in adduction, elbow flexion to 90	Sitting

									degrees, forearm and wrist in neutral position.	
Guerra [196]	2017	Yes	Yes	No	Non-dominant	N/A (single session)	JAMAR, Bodygrip with curve-shaped handle, Bodygrip with straight handle	Second	The shoulder of the non-dominant extremity was adducted and neutrally rotated, with the arm naturally rested by the side of the body. The elbow was flexed to 90 degrees, and the forearm and wrist in neutral position.	Sitting or lying
Haidar [475]	2004	Yes	Yes	Not given	Not clear	Mean of 3 & highest of 3	JAMAR	Adjusted to the individual's hand size	As per ASHT[1] [123] recommendations	Sitting
Hamilton [188]	1992	Yes	Yes	Yes	Non-dominant	Mean of 3	JAMAR and sphygmomanometer	Third	As per ASHT[1] [123] recommendations	Sitting
Hamilton [181]	1994	Yes	Yes	Yes	Right & left	First & highest of 3 & mean of 2 & mean of 3	JAMAR	Second, Third, Fourth	The elbow was at 90 degrees of flexion. No other details provided.	Sitting
Haward [476]	2002	Not clear	Yes	Not given	Dominant & non-dominant	Median of 3	JAMAR	Adjusted to the individual's hand size	Subjects sat with their elbows flexed to 90 degrees, wrist in neutral position and forearm supported on the bench.	Sitting
Hilgenkamp [477]	2012	Yes	Not given	Not given	Not clear	Single measurement	JAMAR	Not stated	As per ASHT[1] [123] recommendations	Sitting

Huang [189]	2011	Not given	Not given	Not given	Affected & contralateral	First & mean of 2 & highest of 2 & mean of 3 & highest of 3	Eval Solosystem dynamometer (Greenleaf Medical, Palo Alto, California, USA)	Not stated	Not stated	Not stated
Irwin [478]	2010	Yes	Yes	Not given	Right	Highest of 3	MAP, baseline dynamometer, vigorimeter	Second	As per ASHT[1] [123] recommendations	Sitting
Jenkins [479]	2017	Yes	Yes	Not given	Not clear	Mean of 3	JAMAR	Not stated	Not stated	Not stated
Kennedy [480]	2010	Yes	Yes	Yes	Right & left	First & mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Khamwong [481]	2010	Not clear	Not clear	Not clear	Non-dominant	Mean of 3	Electronic digital HD	Not stated	As per ASHT[1] [123] recommendations	Sitting
MacDermid [482]	1994	Yes	Yes	Yes	Affected & contralateral	First & mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Maher [483]	2018	Not clear	Yes	Not clear	Right & left	First & mean of 3	Baseline Pneumatic Squeeze Bulb Dynamometer	Not stated	As per ASHT[1] [123] recommendations	Sitting
Mawdsley [484]	2001	Not clear	Yes	Yes	Dominant & non-dominant	First & second & third & highest of 3 & mean of 2 & mean of 3	JAMAR	Second	Shoulder adducted and neutrally rotated, elbow fully extended, and forearm in neutral. The position of the wrist was allowed to vary between 0 and 30 degrees extension and 0 and 15 degrees ulnar deviation.	Sitting
Medina-Mirapeix [485]	2016	Not clear	Not clear	Not clear	Dominant	Mean of 2	Not given	Not stated	Shoulder adducted 0 degrees, elbow flexed 90 degrees, and forearm in neutral position.	Sitting

Niebuhr [486]	1994	Not clear	Yes	Yes	Right & left	Mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Nitschke [193]	1999	Yes	Yes	Yes	Dominant	Mean of 3	JAMAR	Third	Shoulder adducted and neutrally rotated, the elbow flexed to approximately 90 degrees, the forearm in the neutral position, and the wrist between 0 and 30 degrees extension and between 0 and 15 ulnar deviation.	Sitting
Paltamaa [487]	2005	Not clear	Yes	Not clear	Right & left	Highest of 3	JAMAR	Adjusted to the individual's hand size	As per ASHT[1] [123] recommendations	Sitting
Peolsson [488]	2001	Yes	Yes	No	Right & left / Affected & contralateral	Highest of 3	JAMAR	Not stated	Shoulder adducted and in neutral rotation, the elbow flexed at 90 degrees, the lower arm and wrist in neutral position.	Standing
Plant [489]	2016	Not clear	Yes	No	Not clear	Mean of 3	manual & electronic	Second	Shoulder adducted and neutrally rotated, elbow was flexed at 90. The forearm was neutrally rotated with the wrist in neutral deviation.	Sitting
Puthoff [490]	2013	Not clear	Not clear	Not clear	Right & left	Highest of 3	Not specified	Second	Not stated	Sitting

Reddon [187]	1985	Yes	Not clear	Not clear	Dominant & non-dominant	Mean of 10	Stoelting/ Smedley hand dynamometer	Not stated	Not stated	Not stated
Reijnierse [195]	2017	Yes	Not given	Not given	Not clear	N/A (single session)	JAMAR	Adjusted to the individual's hand size	The arms were parallel to their trunk. No other details provided.	Standing
Reuter [184]	2011	Not clear	Yes	Yes	Dominant & non-dominant	Highest of 3	Smedley Hand Dynamometer	Not stated	The elbow was flexed at 90 degree, while the forearm was allowed to rest lightly on the arm of the chair or on the subject's thigh. No other details provided.	Sitting
Savva [491]	2013	Yes	Yes	Yes	Dominant	Mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Savva [33]	2014	Not clear	Yes	Yes	Not clear	Mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Savva [128]	2018	Not clear	Yes	Yes	Affected	Mean of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Schaubert [185]	2005	Not clear	Not clear	Not clear	Right & left	Not clear	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Schreuders [492]	2003	Yes	Yes	Not given	Affected & contralateral	Mean of 3	Lode HG	Second, Fourth	As per ASHT[1] [123] recommendations	Sitting
Segura-Orti [493]	2011	Not given	Yes	Yes	Dominant & non-dominant	Highest of 3	Takei	Not stated	Participants were positioned standing with the elbow extended. No other details provided.	Standing
Shechtman [179]	2003	Yes	Yes	Yes	Right & left	Mean of 3	BTE-Primus Grip Tool, JAMAR	Second	Shoulder adducted and neutrally rotated, elbow flexed at 90 degrees, forearm in midprone	Both sitting and standing

									position, and wrist in neutral.	
Shechtman [494]	2005	Yes	Yes	Yes	Not clear	Mean of 3	DynEx & JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Silva [495]	2019	Yes	Not given	Not given	Right & left	Mean of 3	JAMAR	Adjusted to the individual's hand size	The shoulder was in a neutral position and adducted; the elbow was at 90 degrees flexion and the wrist in a neutral position.	Sitting
Smidt [496]	2002	Yes	Yes	Yes	Dominant & non-dominant	Mean of 3	JAMAR	Adjusted to the individual's hand size	The arm was in 30 degrees of adduction, the elbow in 90 degrees of flexion, and the forearm, wrist, and hand were supported.	Standing
Solari [497]	2008	Yes	Not clear	Not clear	Dominant	Highest of 3	Citec CT 3001, CIT Technics BV	Not stated	Not stated	Sitting
Spijkerman [498]	1991	Not clear	Yes	Yes	Not clear	Mean of 3	Strain-gauge dynamometer	Not stated	Not stated	Sitting
Stephens [499]	1996	Yes	Yes	Yes	Dominant & non-dominant	Mean of 3	Tekdyne hand dynamometer, standard JAMAR, modified JAMAR	Not stated	As per ASHT[1] [123] recommendations	Sitting
Stockton [500]	2011	Not clear	Yes	Not clear	Dominant	Highest of 3	JAMAR	Not stated	The elbow was flexed to 90 degrees and the forearm parallel to the floor. No other details provided.	Sitting
Svensson [501]	2006	Yes	Yes	No	Right & left	Mean of 3	Grippit	Not stated	As per ASHT[1] [123] recommendations	Sitting

Tager [502]	1998	Not clear	Not clear	Not clear	Dominant	Not clear	Not given	Not stated	Not stated (only states that "the measures included in this study represent direct assessments that have been used as part of the standard protocols used for community-based epidemiological studies of older populations")	Not stated
Tan [503]	2001	Not given	Yes	Yes	Dominant	Highest of 3	Takei	Adjusted to the individual's hand size	The elbow was comfortably straight and the wrist in mid-pronation. No other details provided.	Standing
Trippolini [504]	2013	Not clear	Not clear	Not clear	Right & left	Mean of 3	JAMAR	Not stated	The shoulder was adducted without internal or external rotation, the elbow was flexed at approximately 90 degrees and the forearm and wrist were in neutral position.	Sitting
Trutschnigg [505]	2008	Yes	Yes	No	Dominant	Mean of 3	JAMAR & Biodex	Third	As per ASHT[1] [123] recommendations	Sitting
Tsang [506]	2005	Yes	Not given	Not given	Dominant & non-dominant	Mean of 3 & highest of 3	JAMAR	Second	As per ASHT[1] [123] recommendations	Sitting
Tveter [507]	2014	Yes	Not given	Not given	Right & left	Mean of 2	Baseline dynamometer	Second	The arm was alongside the trunk and elbow in 90	Sitting

									degrees. No other info provided.	
Vermeulen [508]	2015	Not clear	Yes	Not clear	Right & left	N/A (single session)	Grip-ball, JAMAR	Second	The participants were instructed to rest their forearm on the arm of the chair with their wrist just over the end of it. No other details provided.	Sitting
Villafane [509]	2015	Yes	Yes	Not clear	Affected & contralateral	Mean of 3	JAMAR	Not stated	The subjects were sitting with the shoulder of tested arm adducted to the side, the elbow flexed at 90 degrees, and the forearm and wrist neutrally positioned.	Sitting
Villafane [510]	2016	Yes	Yes	Not clear	Dominant & non-dominant	Mean of 3	Portable JAMAR	Second	The subjects were sitting with the shoulder of tested arm adducted to the side, the elbow flexed at 90 degrees, and the forearm and wrist neutrally positioned.	Sitting

[1]ASHT (i.e., American Society of Hand Therapists) recommendations: Participants sat on a chair, upright, with their feet flat and hips and knees flexed to approximately 90 degree, shoulders adducted, elbows flexed to 90 degree, forearms in a neutral position, and wrists in 0 to 30 degree of extension.

Appendix C7. Quality assessment of the identified primary studies.

Author	Year	Evidence that the patients were stable in the time between the administration of the tests	Evidence that the time interval between measurements of same session was appropriate	Evidence that the time interval between the sessions was appropriate	Evidence that the measurement condition was the same for the tests	Evidence that the professional administered the test without knowledge of other repeated measurements in the same patient	Evidence that statistical methods used for reliability were appropriate	Evidence that statistical methods used for measurement error were appropriate
Abizanda [449]	2012	Very good	Adequate	N/A	Very good	N/A	Adequate	N/A
Aguiar [450]	2016	Very good	Adequate	Very good	Very good	Very good	Very good	Very good
Alencar [183]	2012	Adequate	Very good	Very good	Very good	Inadequate	Adequate	N/A
Alfonso-Rosa [129]	2014	Adequate	Very good	Very good	Very good	Inadequate	Adequate	Very good
Allen [451]	2011	Very good	Adequate	Very good	Very good	Inadequate	Adequate	N/A
Anumula [452]	2014	Very good	Adequate	Very good	Adequate	Inadequate	Adequate	Very good
Baldwin [453]	2013	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Barden [191]	2012	Doubtful	Inadequate	Doubtful	Very good	Inadequate	Adequate	N/A
Bertrand [454]	2015	Very good	Inadequate	Very good	Very good	Very good	Very good	Very good
Blankevoort [455]	2013	Adequate	Inadequate	Very good	Very good	Very good	Very good	Very good
Bodilsen [456]	2015	Very good	Inadequate	Very good	Adequate	Very good	Very good	Very good
Bohannon [457]	2005	Doubtful	N/A	Adequate	Very good	Very good	Very good	Adequate
Bohannon [458]	2006	Very good	Very good	Very good	Very good	Very good	Very good	Adequate
Bohannon [459]	2011	Adequate	Inadequate	Very good	Very good	Inadequate	Adequate	Very good
Boissy [460]	1999	Adequate	Very good	Very good	Very good	Inadequate	Adequate	Adequate
Brogardh [461]	2015	Very good	Very good	Very good	Very good	Inadequate	Very good	Adequate
Brown [462]	2000	Adequate	Doubtful	Doubtful	Very good	Very good	Adequate	N/A
Buehring [463]	2014	Doubtful	Adequate	Very good	Doubtful	Inadequate	Adequate	N/A

Burnstein [464]	2011	Doubtful	Inadequate	Doubtful	Very good	Inadequate	Very good	Adequate
Carbonell-Baeza [465]	2015	Adequate	Very good	Adequate	Very good	Inadequate	Very good	Very good
Chen [466]	2009	Very good	Very good	Very good	Very good	Inadequate	Very good	Very good
Clifford [113]	2013	Doubtful	Adequate	Doubtful	Very good	Inadequate	Very good	Very good
Coldman [182]	2006	Very good	Adequate	Very good	Very good	Very good	Very good	Adequate
Dunn [186]	1978	Very good	Inadequate	Very good	Very good	Inadequate	Adequate	N/A
Ekstrand [467]	2015	Adequate	Very good	Adequate	Very good	Inadequate	Very good	Very good
Essendrop [468]	2001	Adequate	Very good	Adequate	Very good	Inadequate	Adequate	Adequate
Fox [469]	2014	Adequate	N/A	Adequate	Very good	Inadequate	Adequate	Very good
Gatt [470]	2018	Adequate	Doubtful	N/A	Very good	N/A	Adequate	N/A
Gerhardsson [471]	2014	Adequate	Inadequate	Adequate	Very good	Inadequate	Adequate	Adequate
Gerodimos [472]	2012	Very good	Very good	Very good	Very good	Adequate	Adequate	Very good
Gittings [473]	2016	Very good	Inadequate	N/A	Very good	N/A	Very good	Very good
Gittings [474]	2018	Very good	Inadequate	N/A	Very good	N/A	Adequate	Very good
Guerra [196]	2017	Very good	Very good	N/A	Very good	N/A	Very good	Adequate
Haidar [475]	2004	Very good	Very good	Adequate	Very good	Inadequate	N/A	Very good
Hamilton [188]	1992	Adequate	Very good	Adequate	Very good	Inadequate	Adequate	N/A
Hamilton [181]	1994	Very good	Adequate	Adequate	Very good	Inadequate	Adequate	N/A
Haward [476]	2002	Adequate	Doubtful	Adequate	Very good	Inadequate	Adequate	N/A
Hilgenkamp [477]	2012	Adequate	N/A	Adequate	Very good	Inadequate	Adequate	N/A
Huang [189]	2011	Adequate	Inadequate	Adequate	Very good	Inadequate	Adequate	Very good
Irwin [478]	2010	Adequate	Very good	Adequate	Very good	Adequate	Very good	Adequate
Jenkins [479]	2017	Adequate	Inadequate	Adequate	Very good	Inadequate	Very good	Very good
Kennedy [480]	2010	Very good	Very good	Inadequate	Very good	Inadequate	Adequate	Adequate
Khamwong [481]	2010	Very good	Very good	Very good	Very good	Very good	Very good	Very good
MacDermid [482]	1994	Very good	Adequate	Very good	Very good	Very good	Adequate	N/A
Maher [483]	2018	Adequate	Doubtful	Adequate	Very good	Inadequate	Adequate	N/A
Mawdsley [484]	2001	Very good	Adequate	Very good	Very good	Very good	Very good	Adequate
Medina-	2016	Very good	N/A	Very good	Very good	Very good	Adequate	Adequate

Mirapeix [485]								
Niebuhr [486]	1994	Adequate	Very good	Adequate	Very good	Inadequate	Very good	N/A
Nitschke [193]	1999	Very good	Adequate	Very good	Very good	Doubtful	Adequate	N/A
Paltamaa [487]	2005	Very good	Very good	Very good	Very good	Inadequate	Very good	Very good
Peolsson [488]	2001	Adequate	Inadequate	Adequate	Very good	Very good	Adequate	N/A
Plant [489]	2016	Adequate	Adequate	Adequate	Very good	Inadequate	Very good	N/A
Puthoff [490]	2013	Very good	Inadequate	Very good	Very good	Inadequate	Very good	Very good
Reddon [187]	1985	Adequate	Very good	Adequate	Very good	Inadequate	Adequate	N/A
Reijnierse [195]	2017	Very good	Inadequate	N/A	Adequate	N/A	Very good	Very good
Reuter [184]	2011	Very good	Very good	Very good	Very good	Inadequate	Adequate	N/A
Savva [491]	2013	Very good	Adequate	Very good	Very good	Very good	Very good	Adequate
Savva [33]	2014	Very good	Adequate	Very good	Very good	Very good	Very good	Adequate
Savva [128]	2018	Very good	Adequate	Very good	Very good	Very good	Very good	Adequate
Schaubert [185]	2005	Very good	Very good	Very good	Very good	Very good	Very good	Very good
Schreuders [492]	2003	Adequate	Inadequate	Adequate	Very good	Adequate	Adequate	Very good
Segura-Orti [493]	2011	Adequate	Adequate	Adequate	Very good	Inadequate	Very good	Very good
Shechtman [179]	2003	Very good	Very good	Very good	Very good	Inadequate	Adequate	N/A
Shechtman [494]	2005	Very good	Very good	Very good	Very good	Inadequate	Adequate	N/A
Silva [495]	2019	Very good	Inadequate	Very good	Very good	Very good	Very good	Very good
Smidt [496]	2002	Inadequate	Adequate	Inadequate	Very good	Very good	Very good	Very good
Solari [497]	2008	Adequate	Doubtful	Adequate	Very good	Adequate	Adequate	N/A
Spijkerman [498]	1991	Adequate	Very good	Adequate	Very good	Inadequate	Adequate	Very good
Stephens [499]	1996	Very good	Very good	Very good	Very good	Inadequate	Very good	Adequate
Stockton [500]	2011	Very good	Very good	Very good	Very good	Inadequate	Very good	Very good
Svensson [501]	2006	Adequate	Very good	Adequate	Very good	Inadequate	Very good	Very good
Tager [502]	1998	Very good	Very good	Very good	Very good	Inadequate	Very good	N/A
Tan [503]	2001	Very good	Very good	Very good	Very good	Inadequate	Adequate	N/A
Trippolini [504]	2013	Very good	Inadequate	Very good	Very good	Very good	Very good	Very good
Trutschnigg [505]	2008	Adequate	Inadequate	Adequate	Very good	Inadequate	Adequate	Adequate
Tsang [506]	2005	Very good	Adequate	Very good	Very good	Inadequate	Very good	Adequate

Tveter [507]	2014	Very good	Inadequate	Very good	Very good	Inadequate	Very good	Adequate
Vermeulen [508]	2015	Very good	Inadequate	N/A	Very good	Inadequate	Adequate	Adequate
Villafane [509]	2015	Adequate	Very good	Adequate	Very good	Inadequate	Very good	Very good
Villafane [510]	2016	Adequate	Very good	Adequate	Very good	Inadequate	Very good	Very good