CAN ROUTINELY COLLECTED PRIMARY CARE DATA BE USED TO PREDICT FUTURE RISK OF MORBIDITY AND MORTALITY IN NEWLY-DIAGNOSED TYPE 2 DIABETES MELLITUS?

BY

RONAN RYAN

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

School of Health and Population Sciences
The University of Birmingham
October 2012

ABSTRACT

Background and clinical context: Type 2 diabetes (T2DM) is associated with an increased risk of adverse outcomes over a person's lifetime. Data routinely recorded in general practice electronic patient records could be used to develop risk prediction models to identify those at higher risk and target preventative treatment.

Objective: To develop models to predict the 5-year risk of coronary heart disease (CHD), stroke, chronic kidney disease (CKD), and all-cause mortality following a diagnosis of T2DM.

Methods: Newly diagnosed T2DM patients (1998-2003) registered at a practice contributing data to a large UK general practice database (THIN) were included in the analyses. The risk models included clinical predictors routinely recorded following diabetes diagnosis plus cardiovascular preventative treatments. Missing baseline risk factors were estimated using multilevel regression and imputation. Outcomes were modelled as time-to-event.

Results: 20041 patients diagnosed with T2DM were included. The proportion of variation explained by each model (R^2) was: CHD 0.09; stroke 0.35; CKD 0.34; and mortality 0.58. Hazard ratios for modifiable risks in the mortality model were: current smoking 1.65; blood pressure (high/treated) 1.07; and glycaemic control (HbA_{1C}/%) 1.09 ($p<0.01$ apart from blood pressure). For non-modifiable risks, hazard ratios were: 1.10 age (/year); 1.29 male sex; 1.58 prior CHD; 1.47 prior stroke; and 1.33 prior CKD. Hazard ratios were similar or lower in the morbidity models other than blood pressure (1.80 stroke 1.41-45 CHD/CKD, $p<0.05$). Raised/treated cholesterol was not a consistent risk factor.

Conclusion: The models were predictive, particularly for mortality, and suggest that older, male, smokers, those with poor blood pressure and glycaemic control and those with cardiovascular co-morbidity are at highest risk and should be targeted at the point of diagnosis. The models could be used to highlight such patients and potentially as an educational tool.

ACKNOWLEDGEMENTS

I would like to thank the following people and organisation:

My PhD supervisors Richard McManus and Sue Wilson, and latterly, Tom Marshall for their time and support.

Roger Holder for his support on statistical methodology.

Angela, Katy and Joe for their love and extended patience.

EPIC (now Cegedim Strategic Data Medical Research UK), owner of The Health Improvement Network (THIN) database, for the THIN data used in this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS

ACE	Angiotensin-converting enzyme inhibitors
AHD (codes)	Additional Health Data codes: a set of codes used in the THIN database to record information on patient lifestyle, health factors, numeric observations and laboratory results
AHD (data table)	The THIN database table that contains information on patient lifestyle, health factors and numeric observations and laboratory results
AMR (date)	Acceptable mortality recording date: used in the THIN database to denote the date from which a practice was recording mortality consistently.
ARB	Angiotensin receptor blockers
BMI	Body mass index
BP	Blood pressure
CDF	Cumulative distribution function
CHD	Coronary heart disease
CiPCA	The Consultations in Primary Care Archive: a small regional general practice database
CKD	Chronic kidney disease
CPRD	The Clinical Practice Research Datalink (formerly known as the GPRD)
CVD	Cardiovascular disease
DARTS	Diabetes Audit and Research in Tayside: a disease-specific database
DIN-LINK	The Doctor's Information Network general practice database
eGFR	Estimated glomerular filtration rate
EMIS	Egton Medical Information Systems: a UK company which makes clinical computer systems. Also commonly used as the name of their system. All practices

	that contribute to the QResearch database use the EMIS clinical computer system
EPIC	The Epidemiology and Pharmacology Information Core
exp	Exponential
GMS (contract)	The General Medical Services contract between general practices and primary care organisations for delivering primary care services to local communities
GP	General practice / general practitioner
GP2GP	The national project that allows electronic patient records to be transferred directly from general practice to general practice
GPRD	The General Practice Research Database (now known as the CPRD)
GPs	General practitioners
HbA_{1C}	Haemoglobin A_{1C}
HDL	High density lipoprotein cholesterol
HES	Hospital Episode Statistics
HF	Heart failure
HR	Hazard ratio
ICE	Imputation by chained equations: interchangeable with MICE (multiple imputation by chained equations)
IMD	Index of Multiple Deprivation: an area-based measure of material deprivation
InPS	In Practice Systems Ltd: a UK company which makes clinical computer systems
KM	Kaplan-Meier
LDL	Low density lipoprotein cholesterol
LLD	Lipid-lowering drug
MAR	Missing at random: see table 6.8
MCAR	Missing completely at random: see table 6.8

MED (data table)	The THIN database table that contains all conditions and symptoms entered on the practice computer during consultations between the GP/nurse and patient
MI	Myocardial infarction
MI	Multiple imputation
MODY	Maturity onset diabetes of the young
NHS	The National Health Service
NMAR	Not missing at random: see table 6.8
NPfIT	National Programme for IT in England
OTC	Over-the-counter: drugs which can be bought without a prescription
PAT (data table)	The THIN database table that contains patient demographic data
PCTs	Primary Care Trusts
PRIMIS	National Health Service funded organisation aiming to improve the consistency and completeness of data held on general practice systems
PVI (data table)	The THIN database table that contains postcode-based socioeconomic, ethnicity and environmental indicators for individual patients
QOF	The Quality and Outcomes Framework
QResearch	The QResearch general practice database
R^2	Overall statistic to describe the proportion of variation in the data explained by a model: see section 6.8.2
RCTs	Randomised controlled trials
Read (codes)	A coded thesaurus of clinical terms: the basic means by which clinicians record patient findings and procedures in health and social care IT systems across primary and secondary care in the UK
RSS	Residual sum of squares
S(t)	The survival function

SBP	Systolic blood pressure
SCR	Summary Care Record
SD	Standard deviation
STROBE	An international, collaborative initiative of epidemiologists, methodologists, statisticians, researchers and journal editors involved in the conduct and dissemination of observational studies, with the common aim of STrengthening the Reporting of OBservational studies in Epidemiology
T2DM	Type 2 diabetes mellitus
THE (data table)	The THIN database table that contains practice prescribing data
THIN	The Health Improvement Network
UKPDS	The UK Prospective Diabetes Study
UTS	Up-To-Standard: GPRD term used to describe practices that meet minimum data recording standards
Vision	The name of one of the clinical computer systems made by In Practice Systems
VM	The name of one of the clinical computer systems made by In Practice Systems. Precursor to Vision.
γ	Gamma
λ	Lambda

CHAPTER 1

INTRODUCTION TO TYPE 2 DIABETES AND THE VALUE OF BEING ABLE TO ASSESS INDIVIDUAL RISK OF IMPORTANT OUTCOMES

1.1 Introduction

This chapter introduces the disease of interest in this thesis, type 2 diabetes, and the rationale for developing diabetes-specific multivariate models to predict the risk of several important outcomes in this population.

1.2 Type 2 diabetes

Diabetes mellitus is a metabolic disorder of multiple aetiology characterised by chronic hyperglycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. (1) It may be diagnosed as a result of screening or the investigation of symptoms such as increased thirst, urine volume, weight loss or recurrent infections. The testing process and cutoff values recommended for diagnosis have changed over time: one or more blood glucose measurements and oral glucose tolerance test alone were recommended until 2011, when glycated haemoglobin (HbA_{1C}), a measure of average blood glucose levels over 2-3 months, was included as a method of diagnosis. (2, 3)

The two most common categories of diabetes are type 1 and type 2. Type 2 is more prevalent and is caused by a combination of resistance to insulin action and impaired insulin response. (4) As there may be no overt symptoms after onset, it can be present for a number of years before it is diagnosed, and once diagnosed it is usually present for the remainder of the person's life. (5) It is also relatively common and incidence and prevalence is increasing in countries like the UK. (6-8)

Primary care is central to the management of type 2 diabetes in the UK, and there have been national standards of care for people with diabetes since the early 1990s (9, 10) The public health and clinical importance of this disease in the UK is evident in the range of diabetes related guidelines on its management, the management of related risk factors, and the remuneration offered for good quality management in primary care. (11-16)

Type 2 diabetes, and diabetes in general, is associated with an increased risk of fatal and non-fatal outcomes over the person's lifetime, including damage to the eyes, kidneys, nerves, heart, and blood vessels. (17-22) These outcomes are costly both to individuals and to health services. There are known risk factors associated with these adverse outcomes, including blood glucose control, blood pressure, cholesterol, obesity, and smoking which can be modified by lifestyle changes and medical management. (23-35)

Although nephropathy and subsequent chronic kidney disease is a microvascular complication of diabetes, the main risks considered in this thesis are macrovascular, that is, the increased risk of coronary heart disease, stroke and death that results from atherosclerosis. (17-22) Other diabetic-related microvascular diseases such as retinopathy and neuropathy are not considered here.

1.3 The value of being able to assess risk in a clinical setting

As described in the previous section, some CVD and CKD risks are modifiable or at least the rate of decline in function can be reduced by appropriate treatment. For the most part, these are the same risk factors that influence risk in the general population: smoking; blood pressure, lipid levels and obesity. (36-38) For people with diabetes, there can be additional predictors of risk, including poor blood glucose control. (39) Early intervention that targets these individual factors may reduce the risk of serious outcomes like CVD and CKD by delaying or preventing their occurrence.

It may also be a better approach to assess risk factors in combination than in isolation, that is, to combine them into a single estimate of risk on which to base treatment decisions. There is evidence that blood pressure lowering and statin treatment are beneficial to people with low- and high cardiovascular risk (40-42), and that the benefits of treatment are proportional to risk. (43)

Unlike the individual risk assessment, suggested above, risk prediction models can be used to assess whole populations, for example in a single general practice, quickly and in a standard manner. (38) This cannot be done so easily by individual clinicians, particularly as these risks may not be present at the point of diagnosis of diabetes, but may only become clinically significant at a later date.

When applied to whole populations, the use of risk prediction models can help to allocate resources to those most likely to benefit as the benefits of treatment are proportional to baseline risk. (40-43) These models can stratify the population into groups with a similar level of risk and allow an appropriate intensity of treatment to be targeted to appropriate patients.

1.4 Why would you want to develop models specific to type 2 diabetes?

The main benefit of developing diabetes-specific models is the ability to include predictors, like HbA_{1C}, which are routinely available for this population and are known to have an impact on the risk of important outcomes. (39)This maximizes the applicability of these models to clinical practice. There are two additional reasons why this may be an appropriate approach to model development.

Firstly, the effect of specific risk factors on risk may differ between diabetic and non-diabetic populations. Yudkin and colleagues developed separate models for people with and without diabetes based on data from the Framingham Heart Study. (44) The effect of smoking, BP and cholesterol on CHD risk differed in these two groups, suggesting that models which combine these two groups without including interactions between these risk factors and diabetes may result in models which predict risk in non-diabetic populations better than diabetic populations. Other analyses of UKPDS data, showed that there is an important distinction between age at diagnosis of diabetes and time since diagnosis, and that there is some evidence that diabetic dislipidaemia is quantitatively different from dislipidaemia in the general population. (45)

Secondly, the risk of many outcomes is greater in people with diabetes than in the general population: 2- to 5-fold for stroke. (46) As patients at highest risk are likely to benefit most from intervention, it is therefore important to estimate risk in people with diabetes accurately. (43)

1.5 Summary

This chapter presented the rationale for developing type 2 diabetes-specific risk prediction models. The next chapter describes the previous attempts to develop models that could be used in this population.

CHAPTER 2

PREVIOUS ATTEMPTS TO PREDICT RISK IN PEOPLE WITH TYPE 2 DIABETES

2.1 Introduction

As indicated in the previous chapter, this thesis focuses on the risk of CVD (separately as CHD and stroke), CKD and all-cause mortality following a diagnosis of type 2 diabetes. This chapter introduces previous models which could be used predict these outcomes. These were identified in a rapid review of models published between 1991 and 2012. (47) The models presented for CVD and mortality are specific to type 2 diabetes as a large number of examples were identified: the models for remaining outcome, CKD, predict risk in the general population and make some adjustment for the presence of diabetes as an additional independent risk factor. The features of these models are also described: what predictors they include and how specific these are to diabetes. The last section describes the range of risk factors that should be included in any new models, and the design and methodological limitations of existing models.

2.2 Rapid review of risk models for CVD, CKD and all-cause mortality

Abstract: To identify all papers that presented a CVD (i.e. CHD or stroke), CKD, or all cause mortality prediction model developed in patients with diabetes or that could be applied to individuals with type 2 diabetes.

Methods: Separate online PubMed searches for each outcome of interest (http://www.ncbi.nlm.nih.gov/pubmed). A detailed description of the search and selection process can be found in appendix 2.1. As there was a recent review of CVD prediction models in patients with type 2 diabetes (48), the PubMed search was restricted to the period 1/4/2011-31/12/2012 to identify any additional studies published since this review. The search period for the remaining two outcomes was from 1/1/1991-31/12/2012. In summary, the titles, abstracts and full-text of the publications were reviewed to identify eligible models. The eligibility criteria for prediction models were as follows: (a) The prediction model was either developed in people with diabetes or included diabetes as a predictor. (b) The outcome of the prediction model was CVD or CKD or a CVD/CKD component (i.e., CHD, stroke, end stage renal failure, kidney dialysis, kidney transplant), or all-cause mortality. (c) It presented a specific prediction rule/model with sufficient information on all variables to calculate the CVD, CKD risk or in a different population (beta coefficients of the model or otherwise a scoring system/graph/score card/nomogram was provided). Additional papers were then identified by reviewing the additional articles associated with studies which met the eligibility criteria on the PubMed website.

Results: Four systematic reviews were known prior to running the PubMed searches. (48-51) A total of 12 models that predicted the risk of incident CVD were identified by the rapid review or from these systematic reviews. Ten models which predicted the risk of incident CKD in the wider population were identified, but none that predicted the risk of incident CKD specifically. Six diabetes-specific models that predicted the risk of all-cause mortality were identified. These results and the characteristics of the prediction models are described in more detail in the remainder of this chapter.

2.3 Previous risk models for CVD, CKD and all-cause mortality

This section introduces the existing risk models for CVD CKD, and all-cause mortality that are applicable to people with newly-diagnosed type 2 diabetes. The searches were restricted to models which were developed using people with diabetes, and to incident disease, where possible. This was possible with CVD and all-cause mortality, but not CKD as no diabetes-specific models of incident CKD were available for inclusion. The models identified for each of the outcomes of interest are described separately, below, and their details are listed in tables 2.1 to 2.3.

2.3.1 CVD models: specific outcomes reported and populations included in development

Twelve models that predicted the risk of incident CVD in people with diabetes were identified from the literature and are listed in table 2.1. (44-46, 52-60) Five of these predicted CVD risk as their main or only outcome (52-56), five predicted CHD risk alone or in combination with CVD as separate outcomes (44, 45, 57, 59, 60), and two predicted the risk of stroke with no other outcomes (46, 58). Three were exclusively UK-based (45, 46, 59): two used data from the UKPDS (45, 46) and one from a regional diabetes register in Scotland (59). Of the remaining community-based models: three were derived from US populations (44, 52, 60), two from Hong Kong (57, 58), and one each from New Zealand, Austria and Sweden (54-56). The final model was based on participants in a multi-country drug trial (53). The number of outcomes observed was not reported for the two earliest models (44, 45): the remaining

observed approximately 200 to 500 outcomes (46, 52-60), and two reported more than 1000 outcomes (1482 and 6479) (55, 56). Overall, these CVD models were published at a rate of approximately one every two years, suggesting that there has been an ongoing interest in the prediction of CVD in people with diabetes over the past two decades (1991-2013).

For completeness, four additional models which predict risk in the general population were reviewed to identify how they adjusted for the presence of diabetes in their respective model: Framingham, Assign and QRisk and PROCAM. (61-65) In each of these models diabetes was entered as a single covariate, effectively as a single adjustment to the overall predicted risk. This did not allow the risk of other outcomes to vary between diabetic and non-diabetic people, but assumed that the level of factors like age, blood pressure and cholesterol influenced CVD risk in the same manner in both groups. As described in the last section, this may not be a safe assumption.

2.3.2 CKD models: populations included in development

No previous diabetes-specific models that predicted the risk of incident CKD (CKD Stage 2-5: eGFR<60 mls/min/1.73m^2) were identified from the literature. One model was identified which predicted the risk of later stages of CKD in people with diabetes, but was not included here as it did not also predict earlier CKD stages. (66) Nine models which predicted the risk of incident CKD in the wider population were identified from the literature and are listed in table 2.2.(67-75) One of these also predicted end-stage renal failure. (71) Only one of the models, based on general practices contributing to the QResearch database was UK-based (71): of the remaining community-based models, five were derived from US populations (of

which two were from separate hypertension registries) (67, 69, 72, 74, 75), two were from Holland (68, 70), and a further one each was derived from a Taiwanese population (73). Most models reported between approximately 200 and 2000 observed outcomes (67-70, 73-75), one reported 5236 (72), and the largest, based again on the QResearch general practice database, reported in excess of 25000 (71). Although the literature search included studies from 1991, the earliest study identified was published in 2004. (75) Seven of the nine models were published in the two years between 2010 and 2012 (67-73), demonstrating an increased interest in the prediction of CKD in recent years.

2.3.3 All-cause mortality models: populations included in development

Six diabetes-specific models that predicted the risk of all-cause mortality were identified from the literature. (76-81) These are listed in table 2.3. Three were derived from UK-based community populations (the UKPDS, general practices contributing to the GPRD database and patients referred to a diabetes service in one location in England). (79-81) Of the remaining community-based models, one was from Denmark and one was from Hong Kong. (76, 77) The final model was based on a trial population with high CVD risk. (78) The UKPDS and clinical trial papers did not report the numbers of deaths that it observed: the remaining four models observed either approximately 500 or 2000 deaths (table 2.3). Five of the six models were published since 2010. The first, the UKPDS-based model, was published about six years earlier, in 2004. The lack of diabetes-specific models identified in the intervening years does not suggest that publications have been missed by the literature search. Rather, it appears that all-cause mortality does not receive the same amount of interest as

other diabetes outcomes like CVD. The UKPDS model was just one of a set of outcome models that resulted from this study. (81) The authors of four of the five models since the UKPDS publication were specifically interested in the effect of HbA_{1C} on mortality, rather than developing a model to predict mortality in a clinical setting, and included the other predictors in an attempt to control for confounding. (76-78, 80)

Table 2.1 Previous CVD models (individuals with diabetes)

Lead author (reference)	Year of publication (study dates)	Study population	Outcome	Sample size	No. of events	Predictors in the model
Mukamal (52)	2013 (1989-1999)	Patients with diabetes >=65 years from the CHS cohort study, three USA regions	Incident CVD	782	265	Age, smoking status, systolic BP, total cholesterol, HDL cholesterol, creatinine, oral/insulin treatment, C-reactive protein, LVH on ECG, ankle–brachial index, internal carotid intima–media thickness
Kengne (53)	2011 (2001-2008)	Individuals with type 2 diabetes from 20-country trial (ADVANCE) (perindopril-indapamide), aged 55 years or over	Incident CVD	7168	473	Age at diagnosis, duration of diabetes, sex, pulse pressure, treated hypertension, atrial fibrillation, retinopathy, HbA_{1C}, urinary albumin/creatinine ratio, non-HDL cholesterol
Davis (54)	2010 (1993-1998)	Individuals with type 2 diabetes from cohort study, Australia	Incident CVD	1240	185	Age, sex, prior CVD, albumin:creatinine ratio, HbA1C, HDL cholesterol, ethnicity
Elley (55)	2010 (2000-2008)	Individuals with type 2 diabetes from cohort study (DCS), New Zealand	Incident CVD	36127	6479	Age at diagnosis, diabetes duration, sex, systolic BP, smoking, total:HDL cholesterol ratio, ethnicity, HbA_{1C}, albumin:creatinine ratio
Cederholm (56)	2008 (1998-2003)	Individuals with type 2 diabetes from national register, aged 18-70, Sweden	Incident CVD	11646	1482	Age at diagnosis, diabetes duration, sex, smoking, BMI, HbA_{1C}, systolic BP, antihypertensive drug use lipid-lowering drug use
Yang (57)	2008 (1995-2005)	Individuals with type 2 diabetes from diabetes registry, free of heart failure, Hong Kong	Incident CHD	3521	181	Age, diabetes duration, sex, smoking, eGFR, albumin:creatinine ratio, non-HDL cholesterol, total:HDL cholesterol ratio, HbA1C, Systolic BP
Yang (58)	2007 (1995-2005)	Individuals with type 2 diabetes from diabetes registry, Hong Kong	Incident stroke	3668	190	Age, HbA_{1C}, albumin:creatinine ratio, CHD
Donnan (59)	2006 (1995-2004)	Individuals with type 2 diabetes and complete data from regional register (DARTS), Scotland	Incident CHD	4569	243	Age at diagnosis, duration of diabetes, HbA_{1C}, smoking, sex, systolic BP, treated hypertension, total cholesterol, height
Folsom (60)	2003 (1987-1998)	Individuals with diabetes in cohort study (ARIC), aged 45-64 years, from four communities in USA	Incident CHD	1500	257	Age, race, smoking, total cholesterol, HDL cholesterol, systolic BP, use of antihypertensives, smoking status. BMI, waist:hip ratio, heart rate, physical activity, FEV, Keys score, tobacco pack-years, creatinine, albumin, factor VII, WBC, LVH, carotid IMTfactor VIII, von Willebrand factor
Kothari (46)	2002 (1977-NR)	Individuals with incident type 2 diabetes in cohort study (UKPDS), aged 25-65, without recent or multiple CHD events, UK	Incident stroke	4549	188	Age at diagnosis, duration of diabetes, sex, smoking, systolic BP, total:HDL cholesterol ratio, atrial fibrilation
Stevens (45)	2001 (1977-NR)	Individuals with incident type 2 diabetes in cohort study (UKPDS), aged 25-65, with no recent history of CHD, UK	Incident CHD	4540	NR	Age at diagnosis, sex, ethnicity, smoking, HbA_{1C}, systolic BP, total:HDL cholesterol ratio
Yudkin (44)	1999 (NR)	Individuals with diabetes from 11 cohort studies, USA	Incident CHD	NR (<2138)	NR	Age, sex, smoking, microalbuminuria, total:HDL cholesterol ratio

Table 2.2 Previous CKD models (wider population and predict future risk of, not prevalent CKD)

Lead author (reference)	Year of publication (study dates)	Study population	Outcome	Sample size	Number of events	Predictors in the model
O'Seaghdha (67)	2012 (1995-2008)	Framingham Heart Study participants	Incident CKD	2490	229	Age, diabetes, hypertension, baseline estimated glomerular filtration rate, albuminuria
Alssema (68)	2012 (1989-2005)	Three population-based cohort studies from the Netherlands, aged 28-85 years, no type 2 diabetes, CVD	Incident CKD	6780	22%	Age, smoking, use of antihypertensives, use of lipid-lowering drugs, BMI, waist circumference, family history <65 years of MI/stroke, family history diabetes, history of gestational diabetes
Hanratty (69)	2011 (2000-2007)	Hypertension disease registry at Kaiser Permanente USA	Incident CKD	43,305	5236	Age, gender, race/ethnicity, baseline eGFR, baseline and time-varying SBP, HDL cholesterol, BMI, diabetes, CHD, CVD, heart failure, PVD
Halbesma (70)	2011 (1997-2006)	PREVEND observational cohort study participants, Netherlands	Incident CKD with highest quintile in decline in renal function	6809	272	Baseline eGFR, age, urinary albumin excretion, systolic BP, C-reactive protein, and known hypertension (diabetes was included in an early version of the model)
Hippisley-Cox (71)	2010 (2002-2008)	QResearch UK general practice database	Incident moderate-severe CKD. Incident end-stage kidney disease	1574749	25320	Age, ethnicity, deprivation, smoking, BMI, systolic blood pressure, diabetes, rheumatoid arthritis, cardiovascular disease, treated hypertension, congestive cardiac failure, PVD, NSAID use, family history of kidney disease, systemic lupus erythematosis, kidney stones
Hanratty (72)	2010 (2000-2006)	Hypertension disease registry, Colorado, USA	Incident CKD	10420	429	Age, sex, race/ethnicity, marital status, language, diabetes, vascular disease, heart failure, dyslipidaemia, major psychiatric diagnosis, substance abuse, baseline eGFR
Chien (73)	2010 (2003-2009)	Prospective cohort study, Taiwan	Incident CKD	5168	190	Age, BMI, diastolic BP, type 2 diabetes, stroke, postprandial glucose, HbA_{1C}, proteinuria, uric acid
Kshirsagar (74)	2008 (1987-2003)	Two community-based cohort studies (ARIC, CHS), USA	Incident CKD	14155	1605	Age, sex, race/ethnicity, anaemia, CVD, diabetes, heart failure, PVD, HDL cholesterol
Fox (75)	2004 (1978-2001)	Framingham Heart Study participants	Incident CKD	2585	244	Age, sex, baseline eGFR, BMI, smoking, diabetes, systolic BP, hypertension, hypertension treatment, total cholesterol, HDL cholesterol, impaired fasting glucose

Table 2.3 Previous all-cause mortality models (individuals with diabetes)

Lead author (reference)	Year of publication (study dates)	Study population	Outcome	Sample size	Number of events	Predictors in the model
Xu (76)	2013 (1998-2009)	Diabetic cases from Elderly Health Service cohort study (age>=65 years), Hong Kong	All-cause mortality. CVD-, CHD- and stroke-specific mortality	2137	540	Age, sex, education, smoking, alcohol use, exercise, cardiovascular disease history, BMI, total cholesterol, HbA_{1C}
Skriver (77)	2012 (2001-2005)	Individuals with type 2 diabetes from single region in Denmark	All-cause mortality	17760	1859	Age, sex, diabetes duration, mean annual HbA_{1C} at baseline, CVD, arteriosclerosis, acute complication of diabetes, retinopathy, nephropathy, MI, stroke, neuropathy
Andersson (78)	2012 (2003-2009)	Secondary analysis of overweight/obese/high CVD risk individuals with type 2 diabetes from SCOUT trial, 16 countries	All-cause mortality	7479	NR	Age, sex, randomised treatment assignment (sibutramine), diabetes duration, history of arterial hypertension, history of congestive heart failure, history of cardiovascular disease, history of revascularisation, ethnicity, tobacco use, systolic and diastolic blood pressure, heart rate, HbA_{1C}, BMI, HDL cholesterol, LDL cholesterol, urine albumin/creatinine ratio and use of insulin, metformin, thiazolidinediones and sulfonylureas
Kerr (79)	2011 (1999-2007)	Patients referred to type 2 diabetes service, Bournemouth, UK	All-cause mortality	3781	579	Age, sex, year of diagnosis, HbA_{1C} at 3 months, systolic BP, smoking
Currie (80)	2010 (1986-2008)	GPRD patients with type 2 diabetes, whose treatment was changed to combination therapy or insulin and were aged 50 years or over	All-cause mortality	27965	2035	Age, sex, smoking status, cohort (combination therapy or insulin initiated), HbA_{1C}, mean total cholesterol, LVD, Charlson Index
Clarke (81)	2004 (1977-1989)	Patients from the UK Prospective Diabetes Study, newly diagnosed with type 2 diabetes, aged 25-65 years		3642	NR	Age, sex, smoking status, HbA_{1C}, total:hdl cholesterol ratio, MI, renal failure, amputation

2.3.4 Published systematic reviews of CVD and CKD risk models

Four systematic reviews of models used to predict the risk of CVD and CKD were identified: the CVD reviews included models which were either specific to diabetes or designed for the general population, but included the presence of diabetes as a predictor. (48, 49)

The CKD model reviews had a wider scope than was required for this thesis. They aimed to identify all CKD models applicable to the general population, including those which did not list diabetes as a predictor and those which predicted the presence of undiagnosed CKD as well as the risk of future CKD. (50, 51)

The review led by Echouffo-Tcheugui (51) was more optimistic about model quality than that led by Collins (50): it focused on the outcome measures reported and suggested that the use of predictive models in nephrology was not as well established as it is in other clinical areas. The Collins review was more methodological in its critique of the models and found that they were often developed using inappropriate methods and were generally poorly reported. Collins went on to recommend appropriate approaches to the development and validation of prediction models. These are applicable to all prediction models: basing the models on data of appropriate quality; selecting predictors based on the literature and clinical guidelines; handling missing data and continuous covariates appropriately; internal validation using bootstrapping rather than splitting a 'non-massive' dataset into two halves (and using the second half for validation); and reporting methods and results appropriately. Both of the CKD model reviews highlighted the need to externally validate risk prediction models. Collins also pointed out one QResearch paper as an example of good reporting and mentions the

validation of the two models reported in this paper in a second large general practice database. (71)

However, both internal and external validation using routinely collected electronic patient records rely on the same kind of data that were used to develop these prediction models. The predictors and outcomes recorded in routine general practice and used in model development are likely to be affected by the same recording bias as routine GP records from any other source. These sources are, therefore, too similar to the data used in model development to allow their true predictive ability in a clinical setting to be assessed. A more realistic approach to the validation of these models for use in clinical settings would be a prospective clinical study where the model was used to identify those at high and low risk. The predicted and actual level of risk could be compared using data collected from patients participating in the study. Any under- or over-estimation of risk could be assessed by prospective data collection using linked data from primary and secondary care, death certification and disease-specific registries, as appropriate.

2.4 What factors do the existing models identify as predictive of the risk of CVD, CKD and death?

This section identifies the risk factors that should be included in future models that predict the risk of CVD, CKD and mortality in people with type 2 diabetes based on the covariates included in previous models.

2.4.1 CVD

Twelve models were developed to predict CVD risk in people with diabetes: some are single models which predict overall CVD risk and some are specific to CHD and stroke. (44-46, 52-60) These models are listed in table 2.1. Table 2.4 provides an overview of the variables included in each model. The most common single risk factor was current age/age at diagnosis (12 models): this included duration of diabetes in six of the models. The most common group of risk factors included were blood test results: cholesterol values (10 models); albumin/creatinine/eGFR (9); HbA_{1C} (8); microalbuminuria (1); C-reactive protein (1); and carotid IMT factor VII/ factor VII/Von Willebrand factor and WBC count (all mentioned in a single model). The next most common single risk factor was smoking (9 models), followed by demographic factors: sex (9 models) and ethnicity/race (4). Systolic BP was included in 8 models, with pulse pressure, heart rate and ankle-brachial index included in one model each. BMI, waist:hip ratio and height appeared in four models. Current drug treatments were included in several models: antihypertensives (4 models); oral/insulin treatment (1); and lipid-lowering drugs (1). The last group of risk factors consisted of comorbidities at baseline: atrial

fibrillation (2 models); left ventricular hypertrophy (2); prior CVD event (1); and CHD (1). The last two single risk factors were forced expiratory volume (FEV) and Keys score (lipid content of diet), each of which appeared in a single model.

Table 2.4 Variables included in previous CVD models (individuals with diabetes)

	Mukamal (52)	Kengne (53)	Davis (54)	Elley (55)	Cederholm (56)	Yang (57)	Yang (58)	Donnan (59)	Folsom (60)	Kothari (46)	Stevens (45)	Yudkin (44)
Age, age at diagnosis, diabetes duration	X	X	X	X	X	X	X	X	X	X	X	X
Blood/ urine test results (cholesterol, albumin, creatinine, eGFR, HbA_{1C}, microalbuminuria, C-reactive protein, carotid IMT factor VII/ factor VII/Von Willebrand factor, WBC count)	X	X	X	X	X	X	X	X	X	X	X	X
Smoking	X			X	X	X		X	X	X	X	X
Sex		X	X	X	X	X		X		X	X	X
Ethnicity/race			X	X					X		X	
Systolic BP, pulse pressure, heart rate or ankle-brachial index	X	X		X		X		X	X	X	X	
BMI, waist:hip ratio or height					X			X	X			
Current drug treatment (antihypertensives, oral antidiabetic / insulin, lipid lowering)	X	X			X			X	X			
Comorbidity (atrial fibrillation, left ventricular hypertrophy, prior CVD event, CHD, retinopathy)	X	X	X				X	X		X		
Forced expiratory volume, Keys score, physical activity									X			

2.4.2 CKD

Nine models that predicted the risk of future CKD in the general population are listed in table 2.2. (67-75) Table 2.5 provides an overview of the variables included in each model. These included a covariate for diagnosed diabetes in their development (1 model) or in their final model (8), or for a history of gestational diabetes (1). Six of the models included measured blood pressure or diagnosed hypertension. The most common group of risk factors were blood test results: eGFR (5 models); uric acid/urinary albumin/proteinuria (5); cholesterol values/ diagnosed hypercholesterolaemia (4); HbA_{1C}/glucose/impaired fasting glucose (3); and C-reactive protein (1). Demographic variables were the next most common group of risk factors: age (all 9 models); sex (4); race/ethnicity/language (4); material deprivation (1); and marital status (1). BMI/waist measurement was included in five models and smoking status in three. Prior comorbidity was the next most common group of covariates: CVD, heart failure and PVD appeared in a model on three occasions each; and anaemia, kidney stones, substance abuse, major psychiatric disorder, systemic lupus erythematosis, and rheumatoid arthritis on one occasion each. Drug treatments were also included in several models: antihypertensives (2 models); lipid-lowering drugs (1); and NSAID use (1). Family history was the last group of risk factors: family history of CVD, kidney disease and diabetes appeared on one occasion each.

Table 2.5 Variables included in previous CKD models (wider population and predict future risk of, not prevalent CKD)

	O'Seaghdha (67)	Alssema (68)	Hanratty (69)	Halbesma (70)	Hippisley-Cox (71)	Hanratty (72)	Chien (73)	Kshirsagar (74)	Fox (75)
Diabetes as comorbidity or history of gestational diabetes	X	X	X		X	X	X	X	X
Blood pressure or diagnosed hypertension	X		X	X	X		X		X
Blood/ urine test results (eGFR, uric acid, urinary albumin, proteinuria, cholesterol values or diagnosed hypercholesterolaemia, HbA_{1C}, blood glucose, impaired fasting glucose, C-reactive protein)	X		X	X		X	X	X	X
Age	X	X	X	X	X	X	X	X	X
Sex			X			X		X	X
Race, ethnicity or language			X		X	X		X	
Material deprivation					X				
Marital status						X			
BMI or waist measurement		X	X		X		X		X
Smoking status		X			X				X
Comorbidity (CVD, heart failure, PVD, anaemia, kidney stones, substance abuse, major psychiatric disorder, systemic lupus erythematosis, rheumatoid arthritis)			X		X	X	X	X	
Current drug treatment (antihypertensive, lipid-lowering, NSAID)		X			X				X
Family history (CVD, kidney disease, diabetes)		X			X				

2.4.3 Mortality

Six models that predicted the mortality risk in the people with diabetes are listed in table 2.2. (76-81) Table 2.6 provides an overview of the variables included in each model. Demographic variables were the most common group of risk factors included in these models: sex (6 models); age (6); education (1); and ethnicity (1). All models include CVD as a comorbidity: other comorbidities were less commonly included: nephropathy (3 models); heart failure (2); amputation, retinopathy, history of arterial hypertension, arteriosclerosis, MI, stroke and revascularisation (1 model each). One model included a single covariate covering all acute diabetes-related complications and an overall comorbidity score (the Charlson Comorbidity Index). Diabetes-specific risk factors were the next most common group: HbA_{1C} (all 6 models); diabetes treatment type (3); diabetes duration (2); and year of diabetes diagnosis (1). Smoking status was the next most common single risk factor (5 models): BMI, another lifestyle-related risk factor, was included in just two models, and exercise and alcohol use were mentioned together in one further model. Cholesterol level was included in four models and blood pressure in two, and a history of arterial hypertension was included in one model. There was only one further blood result included in these models: urine albumin/creatinine ratio, which appeared in one model.

Table 2.6 Variables included in previous all-cause mortality models (individuals with diabetes)

	Xu (76)	Skriver (77)	Andersson (78)	Kerr (79)	Currie (80)	Clarke (81)
Sex	X	X	X	X	X	X
Age	X	X	X	X	X	X
Education	X					
Ethnicity			X			
Comorbidities (CVD, nephropathy, heart failure, amputation, retinopathy, history of arterial hypertension, arteriosclerosis, MI, stroke, revascularisation, any acute diabetes-related comorbidity, Charlson Comorbidity Index)	X	X	X		X	X
HbA_{1C}	X	X	X	X	X	X
Diabetes treatment type			X		X	X
Diabetes duration, year of diabetes diagnosis		X	X	X		
Smoking	X		X	X	X	X
BMI	X		X			
Exercise, alcohol use	X					
Cholesterol level	X		X		X	X
BP, heart rate			X	X		
Urine albumin/creatinine ratio			X			

2.5 What range of risk factors should be considered for use in future models that predict the risk of CVD, CKD and death in people with type 2 diabetes?

The three groups of models identified in tables 2.1 to 2.3 can be used to estimate the risk of CVD, future CKD and mortality in people with diabetes. They vary in complexity, containing from four to 22 covariates. (58, 60, 78) . Some include risk factors which are not routinely recorded in primary care, even after a diagnosis of diabetes: lipid content of diet (which is also difficult to measure accurately) and von Willebrand factor, for example. (60) Others exclude risk factors which would be routinely recorded in primary care, particularly following a diagnosis of diabetes, and which are known to be associated with CVD, CKD and mortality risk: sex; systolic blood pressure; smoking status, and BMI, for example. (54, 58, 79)

The variables included in the models above suggest that there is a wide set of potentially significant predictors that should be considered for inclusion in any new predictive model for these outcomes. These are the demographic, lifestyle, comorbidity, biochemical, treatment, and clinically observed risk factors that may be associated with each outcome. The minimum specification for any new model should, therefore, include relevant individual predictors from each group in this set and should be adequately powered to detect the effect of changes in their level on the outcome of interest.

2.6 Summary

This chapter introduced existing models which predicted the risk of CVD, CKD and mortality which can be used following a diagnosis of type 2 diabetes. The CVD models which were developed for use with the general population did not include HbA_{1C}, an important risk factor for the outcomes of interest, and by including the presence of diabetes as a single term in their models, without interaction terms with other covariates, made a strong assumption that the effect of diabetes is independent of other risk factors such as blood pressure, cholesterol and comorbidities at baseline. This would tend to limit their ability to accurately estimate risk in type 2 diabetes and their utility in UK primary care. Many of the models specific to diabetes have attempted to include data which would be routinely available in UK primary care and could, therefore, be used in clinical practice in this setting. However, their utility in this setting is limited because they: excluded cases diagnosed at older ages and were based on BP- and cholesterol-untreated populations; included only cases with complete risk factors; and may not have had the power to detect the effect of important risk factors.

CHAPTER 3

REASONS WHY YOU MIGHT WANT TO USE PRIMARY CARE DATA AS OPPOSED TO OTHER DATA SOURCES TO DEVELOP A PREDICTIVE MODEL

3.1 Introduction

The last chapter introduced previous models that predicted the risk of CVD, CKD and death in people with diabetes. These models were derived from a variety of data sources including clinical trials, single and aggregated cohort studies, disease registers, and secondary and primary care. These data were also from a number of countries and refer to events that occurred over the past three decades. Few of these studies, taken individually, would therefore be regarded as automatically valid for use in the current UK population without further evaluation in this population. A hierarchy of evidence for statistical prediction models is suggested below. It emphasises the importance of the representativeness of the population used for development and later external validation of models, rather than the meta-analyses and systematic reviews used to assess other types of research question.

The utility of the possible sources of data for the prediction models of interest are discussed: the strengths and weaknesses of the alternatives and the selected data source (a large primary care database) are then compared.

3.2 Hierarchy of evidence for statistical prediction models

This section suggests a hierarchy of evidence which can be used to assess the utility of various sources of data for use in the development of the statistical prediction models of interest in this study. These models will predict the risk of CHD, stroke, CKD and all-cause mortality in people with newly diagnosed type 2 diabetes, and are intended for use in the UK general practice population.

The hierarchy of evidence for models which are intended to predict clinical outcomes differs from that required for other types of research question (e.g. the effectiveness of an intervention). (82) Merlin and colleagues did not specifically include statistical prediction models like those developed in the current study, but did provide a hierarchy for studies which aim to identify prognostic factors for disease outcomes. (82) They placed systematic reviews of prospective cohort studies at the top of the hierarchy, followed by all-or-none studies (a rare situation where all or none of the people with a risk factor experience the outcome of interest). This was followed in their hierarchy by secondary analysis of RCT data to identify prognostic factors, then retrospective cohort studies (like the current study), and finally case series or cohort studies of people at different stages of a disease.

This hierarchy may be appropriate for studies which aim to identify common prognostic factors across a range of clinical settings, countries and time periods, but not those like the current study which aim to combine predictive factors into a single statistical model for use in a specific population. A more appropriate hierarchy for a statistical prediction model (based on the systematic review of CKD prediction models by Collins and colleagues (50)) might be:

1) analysis of large population representative of the population in which the model is to be applied with validation (in the population in which the model is to be applied);
2) analysis of small population representative of the population in which the model is to be applied (with validation);
3) analysis of a population less representative of the population in which the model is to be applied (with validation); and
4) analysis of a population different from the population in which the model is to be applied (with or without validation).

The most important features in this hierarchy are the use of a population that is representative of the target population for model development; the inclusion (for model development) of a population large enough to provide precise estimates of the value of each of the predictors of interest; and the additional step of model validation in the target population.

The next section describes the utility of the available sources of data for use in the development of predictive models which are intended for use in current UK general practice.

3.3 The utility of prospectively collected data compared with routine data

The possible sources of data for the development of the predictive models of interest in this study can be divided into two broad types: prospectively collected study data from trials and bespoke cohorts, and routine data from primary and secondary care and from disease registers. (82) The utility of these sources can be assessed by the availability and accuracy of outcome and predictor data, the features of the population that they cover (representativeness, duration of follow-up) and cost.

Outcome ascertainment: The outcomes of interest in this study were CVD (CHD and stroke), CKD and all-cause mortality. Studies which collect data prospectively can be designed to ensure that outcomes such as CVD and CKD are collected in a consistent and complete manner. This is unlike routine data sources which rely on these kinds of outcomes being ascertained and recorded in a consistent manner by a potentially large set of clinicians, administrators or clinical coders. Mortality (the fact and date of death, rather than cause), however, may be better recorded in routine data, particularly in primary care because of the link to payment [section 4.6].

Completeness and accuracy of predictor variables: The predictors of interest in this study were demographic (age, sex, material deprivation), comorbidities (prior CVD and CKD), clinical measurements (BMI, BP, cholesterol, eGFR) and drug treatments (BP-lowering and lipid-lowering) as the predictive models were intended for use in UK general practice where these data are routinely recorded for all patients or as part of clinical care following a diagnosis of type 2 diabetes. (39, 83, 84) Other routine data sources (secondary care and diabetes registers) may not collect this full range of predictors (e.g. secondary care may only

collect data which are relevant to the reason for referral or hospital admission: diabetes registers may not have data on comorbidities at diabetes diagnosis). Unlike prospective studies which can arrange for these predictors to be assessed at the same time points and at relatively fixed intervals, these same predictors may be missing (not routinely recorded or not measured for a particular individual), or recorded at different intervals in routine data sources [section 7.5]. Further, the accuracy of measurement may also be greater in prospectively collected study data where measurement protocols can be standardised across study sites and individual investigators. Routinely recorded data, like BP or weight and height, may not have this level of consistency [section 6.6.6].

Population: The target population for the clinical predictions models in this study was current (numbering approximately 10000) UK general practices. (85) Both prospectively collected study data and routine data sources may not be representative of this population: prospectively collected study data may not be representative if they are based on data from a small set of study sites and secondary care data may only reflect those who were seen in hospital for the management of their diabetes or complications arising from it. Given the number of practices in the UK, even large primary care databases comprising of hundreds of practices may not necessarily be representative of all types of practice (e.g. single-handed), all geographical regions, and deprivation. (86)

Cost versus sample size: In principal prospective studies can recruit similar numbers of cases for inclusion in prediction models as studies which use routine data. (87) The limiting factor is cost: the cost of data collection is relatively low with routine data (THIN contributing practices are provided training on the use of their practice software and some feedback on data quality), whereas prospective studies have an ongoing cost associated with collection of predictor data and outcomes for the entire duration of the study.

The remainder of this section focuses on specific examples of each type of data (prospective trial/cohort study data and routine data) which could be used to develop the prediction models of interest in this study. Examples are used from previous prediction models to illustrate the strengths and weakness of each source. Lastly, the strengths and weaknesses of large primary care databases (the source of data selected for this study) are discussed.

Prospectively collected trial/cohort data: The UKPDS (88) was a data source for three of the models identified in the previous chapter (45, 46, 81). The design of this series of studies demonstrates how trial and prospective cohort data in general may not be a suitable source for the predictive models of interest in this thesis.

- **Long interval between study start and model development/publication:** Diabetic complications can take many years to develop (45, 89) : this can introduce a delay of many years before a sufficient number of outcomes to power a multivariate prediction model have been observed. Similarly, if the intention is to develop a model which can predict the 5- or 10-year risk of an outcome, then a proportion of the study participants must be observed for close to 5 or 10 years.
- **Representativeness:** These populations are not necessarily representative of the population which is the target for a predictive model. These differences may include the duration of diabetes, age at diagnosis, health status, case definition and risk management. (45)
    - The UKPDS trial population was limited to people aged 25-65 at diabetes diagnosis: this excluded approximately half of the incident type 2 diabetic population who were over 65 years at diagnosis. (6)
    - It also excluded people if they had recent or multiple CHD events: data from the cohort included in this thesis suggests that over 20% of UK primary care patients

have a history of CHD or stroke at diagnosis of type 2 diabetes (table 7.5). The UKPDS prediction models could not, therefore, reliably be used to predict risk in this group of patients.

- The trial also recruited participants from 1977 to 1991, before statins, which reduce the risk of CVD when used in primary prevention, were in widespread use (fewer than 7000 general practice patients in England were prescribed a statin in 1991 (90)), and did not include any adjustment for exposure to this or other drugs which can reduce the risk of CVD, in particular blood-pressure lowering drugs and antiplatelet agents (45, 46, 81).

Routine data: Routine data refers to data which have been collected (prospectively) for other purposes (usually as part of clinical care) and which are later analysed to answer a research question. (91) The possible sources of data for predictive models considered here are secondary care data, diabetes registers and primary care data.

- **Secondary care data:** Type 2 diabetes mellitus is usually diagnosed and managed in primary care in the UK, rather than secondary care. (92-94) Secondary care data on patients with diabetes is therefore likely to be restricted to patients who were referred to a diabetologist for a particular reason or admitted as an inpatient, and the data items collected in secondary care are likely to be relevant to the reason for referral or admission. (95) Other secondary care models, like that produced by Kerr (based on data from local secondary care led diabetes services data alone) may not reflect the general practice population which would be the target for diabetes-specific predictive models. (79)
- **Diabetes registers:** Diabetes registers, an example of a disease-specific register, cover only relatively small geographic areas in the UK, unlike cancer registries which have national coverage and are relatively few in number. (96, 97) The relationship between the

risk factors and outcomes may differ from other areas. (86) The health services in these areas may also differ from those provided nationally: the presence of a register may result in greater contact between primary and secondary services, and the linked registry data may be used to improve patient outcomes as well as for research. (98) These may limit the representativeness of such diabetes registers as data sources for models which are intended for use across the national population. (59)

- **General practice data:** If predictive models for CVD, CKD and mortality are to be used in a UK primary care setting by GPs, then there are particular advantages to using primary care patient records as the main or only source of data: these range from the availability of data on risk factors, the applicability of these data to the current population, where the models are most likely to be used, and the relevance of the risk factors to clinicians working in primary care. (38) There are also weaknesses to these data, only some of which can be addressed using appropriate statistical methods. (99) On balance, compared with the alternative sources of data and their own strengths and weaknesses, routine records from primary care appears to be an acceptable source of data for the prediction models of interest in this study. They have specific advantages over alternative sources

  Advantages:

  - Primary care is central to the management of type 2 diabetes. (92-94)
  - There is an increased level of contact and risk factor recording in primary care following diagnosis. (12, 13, 84)
  - This increased monitoring and management can provide suitable data for prognostic models. (12, 13, 37, 39, 89, 100)
  - Retrospective cohort studies can be carried out in primary care databases which include in excess of 500 practices from the UK. (64, 71, 101, 102) This avoids the

need for primary data collection from individual practices, and reduces the time taken to produce these models. This is particularly relevant for the current study which is taking place in the context of a PhD.

- The effect of new risk factors can be explored if they are routinely recorded in general practice: these patient records contain a wider range of health-related information than that collected by disease-specific registers and hospital-based services. (96, 103) Primary data collected from general practices or other sources are usually restricted to a set of predictors identified as relevant at the outset. (104, 105) Some primary care database owners provide researchers with the full electronic patient record for cases of interest (THIN and CPRD, but not QResearch): this allows researchers using these data sources to assess the value of additional predictors or explain any unexpected results using additional clinical data about each case.
- The data used to develop these models can be more recent, increasing their applicability to the current population with diabetes. This is because the three main GP databases receive regular data updates from participating general practices. (106-108) In the case of THIN, the gap between data extraction and data being ready for research use can be 5 months. (107) The data provided by the database owner for this study is no longer current (it dates from 2005), but the same modelling can be rerun in more recent data prior to publication to ensure the relevance of the prediction models and take advantage of linkage with secondary care sources and any improvements in the recording of predictors caused by the introduction of QOF. (63, 109)

- Clinicians will make decisions based on the data they have collected in routine primary care. (39)

Disadvantages:

- Large general practice databases may be representative of the wider population in terms of age and sex, but their representativeness may vary by region, and they may not reflect the full range of practices in the UK (each database contains approximately 500 practices (106-108): there are over 10000 practices in the UK (85)). However, there are currently no other larger sources of primary care data, so these are the best available source of primary care data for prediction model development. Models derived from these sources can also be validated and revised using data from dissimilar sets of practices to ensure their representativeness. (64, 71, 109, 110)
- There can be wide variability between general practices in the coding and recording of the data needed to identify cases, outcomes and risk factors. (111) The introduction of QOF and incentives for recording the process of care after a diagnosis of diabetes will have led to more consistent recording of some of the clinical values of interest: this will benefit predictive models which use more recent data for development. (12) Variability in the coding of cardiovascular outcomes may lead to models which over or underestimate the risk of these outcomes. (112)
- Clinical predictors such as cholesterol levels can be missing or not recorded for some cases close to the baseline (diabetes diagnosis) for the clinical prediction models of interest [section 6.6.6]. The exclusion of cases with incomplete risk factor data might lead to biased results or models which are not representative of

the population of interest [section 6.7.5]. (99, 113) However there are appropriate statistical methods for dealing with these issues, such as the estimation of baseline values and multiple imputation [sections 6.6.7 and 6.7.5]. Multiple imputation of missing clinical measurements has been used in other clinical prediction models which are in current use in general practice and are therefore likely to be an acceptable method for dealing with missing data in this study. (64, 71)

3.4 Summary

Although there are alternative sources of data, routine data derived from large primary care databases appear to be an appropriate source of data to generate predictive models intended for use in UK primary care. The next chapter discusses the kind of information that UK primary care and large primary care databases contain with respect to diabetes and related outcomes.

CHAPTER 4

INTRODUCTION TO PRIMARY CARE ELECTRONIC PATIENT RECORDS AND TO LARGE PRIMARY CARE DATABASES

4.1 Introduction

This chapter introduces the electronic patient records held by primary care in the UK and the large primary care databases that are derived from them. The purpose here is to describe how practices use their computer systems and the kinds of data that are entered, particularly with reference to the diagnosis and management of type 2 diabetes. The chapter then goes on to describe what data are available from the three main primary care databases currently operating in the UK. It ends by introducing published evidence on the validity of the diagnoses of interest in this thesis, namely diabetes, CVD and CKD, and their implications for the identification of primary care patients with these diagnoses.

4.2 How practices use their systems and the scope of data recorded by practices

Practices with clinical computer systems typically use them in place of paper records during patient consultations. (12, 114) These systems allow the practice staff to review the details of each patient's history and care, and to store new information on symptoms, diagnoses, test results, and prescribing. (103) They can also provide templates (allowing easy data entry) related to the periodic review of specific chronic conditions, e.g. diabetes. These serve as a reminder of the information that needs to be gathered during the consultation and past events related to that condition. (114)

The extent to which individual practices make use of all the features of their system probably varies from individual clinician to individual clinician and, therefore, from practice to practice, depending on the level of experience of the individual members of staff. (115) It may not, therefore, be used to its full extent for some time after its installation in a practice.

4.3 What data are recorded specifically about diabetes, its management and diabetes-related outcomes and when are they recorded?

The electronic patient records of patients with type 2 diabetes can also contain a record of the diagnosis and management of their diabetes, and the management of any associated risk factors, including relevant outcomes. (6, 12) If the diabetes is diagnosed after the patient registered at their practice, this can include information about the diagnosis itself: the date of diagnosis, blood glucose control (HbA_{1C}) at diagnosis, and the results of any diagnostic tests. (6, 7) Other assessments may be carried out and recorded at this point in order to manage risk factors for microvascular and macrovascular complications of diabetes. (39) For CVD risk, for example, this would involve measuring blood pressure and BMI; identifying the patient's smoking status; and requesting a blood test to measure cholesterol levels. (39) The initiation of drugs to manage diabetes and treat high levels of these risk factors, and subsequent prescriptions issued, are also recorded. (103) Blood glucose control, and the levels and management of other risk factors may be recorded periodically thereafter (annually or more frequently, if required), until the patient leaves their practice or dies. (83, 84, 89)

4.4 What are large primary care databases?

Large GP databases are collections of electronic patient records extracted from individual general practices and have been used for a variety of research studies. (106, 107) The electronic records of individual patients consist of coded data [section 6.2 and table 6.1], and in the case of some databases, anonymised free-text recorded by the practice while the patient was registered with them. (103, 116) The records of patients who have died or left each practice are also available to researchers, in addition to those who are still registered. There are currently three large GP databases in the UK which have been operating for a number of years: the General Practice Research database (GPRD), The Health Improvement Network (THIN), both based in London, and QResearch, based at the University of Nottingham. (106-108) Each contains records from several hundred general practices. There are other smaller GP databases, such as DIN-LINK, and smaller regional databases such as the Consultations in Primary Care Archive (CiPCA) at the University of Keele (which is a subset of QResearch), but these are less widely used and are therefore not of primary concern to this thesis. (117, 118) Any data or methodological issues arising from the use of these smaller databases are likely to be similar to those arising in larger databases, and so will not be addressed separately.

4.5 Which parts of the primary care electronic patient record are made available to researchers using large primary care databases?

Typically, only data which are coded (i.e. Read coded symptom/diagnoses/process of care, and drugs prescribed) or are in numeric format (numeric observations, drug quantity prescribed) are extracted from GP systems and made available for use to researchers. (103) This is intended to preserve the anonymity of patients and practices, but as a consequence, some information which is recorded by clinical staff and available to them when they view a patient's record, is not accessible to researchers.

This includes any historical paper records received from practices where the patient was registered in the past, letters received from secondary care (often stored electronically as scanned files in the clinical computer system) and any free-text comments entered into the computer system during patient consultation. (103) These free-text comments can contain detail that is not coded by the practice (119-121) (personal experience of unanonymised free-text while working as supplier of GPRD data). They often include additional observations or context explaining the significance of any diagnoses, plans for further investigations or treatments, and information gained from specialists, hospital discharge letters and results of diagnostic tests). (122-124) By its nature, free-text is a quicker and more flexible way for practices to enter data, and may provide more detail from consultations than is typically coded. (119) This can limit the ability of researchers to capture important events, like cause of death, and may lead to an underestimate of the frequency of important outcomes, for example myocardial infarction. (112)

4.6 The validity of primary care diagnoses and the recording of death

Three systematic reviews on the validity of primary care data were identified by a literature search. (111, 125, 126) These reviewed studies which validated diagnoses recorded in one of the large UK general practice databases using: additional data recorded in the database; questionnaires sent to general practices; and comparisons with rates from external sources. With respect to the diagnoses of interest in this thesis, one of the reviews suggests that prevalent diabetes is well recorded (positive predictive value, PPV, over 90%), as is CVD (PPV > 90%). (111) None of these reviews reported finding evidence on the validity of CKD diagnoses [section 6.5.3]. Another of these reviews aggregated the individual study findings by broad disease group and found that a median 88% of cases with an endocrine, nutritional and metabolic diagnosis (e.g. diabetes) could be confirmed, and 85% of cardiovascular system diagnoses (e.g. CHD and stroke) could be confirmed. (126) The range for the proportion of diagnoses confirmed in the individual studies for the two groups above ranged from 50%-100%. (126)

One 2005 paper reviewed the electronic patient records of 12 practices and found that less than 4% of patients with a recorded estimate of kidney function (eGFR) in the range for chronic kidney disease had a Read code that indicated that they had CKD. (122) This suggests that any attempt to identify CKD in primary care, especially in the period covered by the study (1998-2003: prior to the introduction of CKD as a QOF domain in 2006), should include estimated GFR in addition to Read coded records to identify patients with CKD.

One study, based on approximately 500 general practices contributing to the QResearch database, estimated that about 1% of the UK population had biochemical evidence of diabetes

in their general practice, but were either undiagnosed or not recorded (i.e. coded) as having diabetes. (127) Two more recent and related studies on the validity of diabetes diagnoses recorded in primary care suggest that approximately 85-90% of practice patients with one of a wide set of diagnostic Read codes for diabetes are true diabetes cases. (128, 129) One issue highlighted by them, and relevant to this thesis, is the miscoding of people with type 2 diabetes using a code for type 1 diabetes. This suggests that any study seeking to identify general practice patients with type 2 diabetes should include diagnostic codes for type 1 diabetes.

Another study, carried out by this author, reviewed the unanonymised records of GPRD patients who had a diagnostic code for diabetes to develop a robust method for identifying cases of type 1 and type 2 diabetes. (7) We concluded that our case definition for prevalent diabetes mellitus should include a wide set of codes which specified type 1 and 2 diabetes and unspecified diabetes mellitus, and exclude codes that specified other types of diabetes (gestational, drug-induced, and diabetes due to haemochromotosis). In addition we found that it would be necessary to exclude patients whose only mention of diabetes was within one year of a pregnancy in order to avoid including women with gestational diabetes only, and those who had a code for cystic fibrosis at any time, in order to avoid including non-type 1 and non-type 2 cases of diabetes. This definition was validated in a subsequent study which identified incident cases of type 1 and type 2 diabetes, again using unanonymised GPRD records. (6) A stratified sample of 143 potential incident cases was reviewed by hand. Of these 12 (8%) had a free-text comment that indicated a new diagnosis and 115 (80%) had no evidence of diabetes monitoring or treatment prior to their first diagnostic code for diabetes and 13 (9%) were possibly or probably diagnosed in an earlier year. The remaining 3 cases (2%) had evidence that they were screened or had ‘borderline’ diabetes. The results of this validation

suggested that its method for identifying new cases of diabetes mellitus would include 89% who were newly diagnosed cases, 9% who were prevalent cases, and a further 2% who did not have diabetes. In order to decrease the risk of including patients without diabetes, we subsequently decided (in that study) to exclude patients with codes for other specified types of diabetes, e.g. diabetes due to haemochromatosis or malnutrition, and neonatal, secondary or 'latent' diabetes and review by hand the records of all potential cases below the age of 25 who did not have evidence of diabetes-specific drug treatment. (6, 7)

Death is an outcome of interest in this thesis and many studies, but as it is not a diagnosis was not included in the above systematic reviews. (111, 125, 126) A comparison of mortality in one large primary care database with national England & Wales data in 2001 found that overall GPRD mortality rates were within 5% of national rates, that cause of death could only be identified in 92% of their sample, and that cardiovascular deaths were probably underestimated in the available GP data (GPRD 33% of deaths; England & Wales 40%) (130). Therefore, it appears that the accuracy and completeness of cause of death available from these databases may not be sufficient for research use. However, the relative (all-cause) mortality of people with incident type 2 diabetes has specifically been addressed in a high impact journal, suggesting that the level of recording of this outcome in this population is acceptable for research use. (131)

4.7 Summary

This chapter described the diabetes-related data that are available from primary care and large primary care databases. The data made available to researchers through these databases contain the coded portion of the full electronic patient record and typically exclude free-text that might confirm or refute the diagnoses and outcomes of interest in this thesis. However the validity of these diagnoses are likely to be sufficient to answer the research question posed in the next chapter, if appropriate definitions of incident diabetes are used and if mortality is restricted to all-cause rather than cause-specific deaths.

CHAPTER 5 RESEARCH QUESTION

5.1 Introduction

This thesis will use a large UK general practice database to carry out a study on the epidemiology of type 2 diabetes mellitus. It will develop four separate statistical models which will predict the risk of coronary heart disease (CHD), stroke, chronic kidney disease (CKD), and all-cause mortality in the five years following diabetes diagnosis.

5.2 Research question

Can routinely collected primary care clinical data be used to predict future risk of CHD, stroke, CKD and/or all cause mortality in people with newly-diagnosed type 2 diabetes mellitus?

CHAPTER 6 METHODS

6.1. Introduction

This chapter describes the methods used to develop the risk prediction models for CHD, stroke, CKD and all-cause mortality in newly diagnosed patients with type 2 diabetes. It begins with an introduction to the data source, the THIN primary care database. It then continues on to describe the criteria used to identify eligible practices and cases, the definitions of the outcomes of interest, and the methods used to identify baseline clinical characteristics and to estimate baseline clinical measurements. The development of the risk prediction models themselves is then described, including the approach to handling missing data. Finally, the methods used to assess the external and internal validity of the models are described.

6.2. Data source

This study used electronic patient records from a single large UK primary care database, The Health Improvement Network database (THIN). (107) THIN began as a collaboration between In Practice Systems Ltd (InPS), the makers of Vision clinical software, and EPIC, the Epidemiology and Pharmacology Information Core. It began collecting data from UK general practices in 2002 (first collecting all available historical data from each practice and then by collecting periodic updates every few months). In common with similar large research databases (CPRD (formerly known as the GPRD) and QResearch), it maintains a database of anonymised patient records from about 500 voluntary participating practices which use a single clinical software system. (107)

All practices that contribute to THIN use Vision which is a self contained clinical software system. (107) At the time of writing, approximately 1800 general practices in the UK use it to store and access the clinical records of their patients. It was gradually introduced from 1994 to replace the earlier VM practice software. In addition to storing clinical information entered by the practices, it is also used to manage scanned paper records from other sources such as letters from hospital, patient appointments, and to support clinical audit within the practices. Although other practices may use software from different suppliers, and patient data may be stored centrally by the system supplier, the use of the software, the way patient information is stored, and the functions provided by the systems are essentially the same whichever system is used. (114)

In return for participation, all practices receive regular feedback on the quality and completeness of their data, summary reports which can be used for external and internal

audits, and free training on the use of their practice software or a payment based on their practice list size. (107)

For the first ever collection, anonymised, coded clinical information is extracted from a tape backup of each practice system by InPS. (107) Practices then have the choice to continue with this extraction method, or, more recently, to install automated collection software on their system for each subsequent collection. These records are then passed to EPIC by InPS where they are checked and added to the existing THIN database. As practices do not make use of area-based data an additional automated matching exercise is carried out by InPS and participating practices periodically to link individual patients to indices of deprivation and environmental data such as air quality using their postcode. These data are then pseudoanonymised before transfer from each practice and passed to EPIC for inclusion in the database. Patients with invalid or missing postcodes cannot be linked to these external datasets, so this additional information will be missing from the THIN database. (103)

A set of consistency checks is run by EPIC on each set of new data. (103) A flagging field is added to each demographic, clinical and prescribing event to indicate if it passed a check. In the case of demographic data, for example, the fields would show if a patient's age lies outside an acceptable range, or if the recorded date for when they left the practice was before the date they first registered at the practice. However, as with other GP database suppliers, there is no attempt made to validate the clinical data in any other way that might be important for researchers: they do not apply any checks to ensure that clinical measurements fall within an acceptable range or identify implausible events for individual patients (e.g. women with a code for prostate cancer). (103, 106, 108) However, some clinical system suppliers do incorporate range checks for numerical data at the time the data are entered at the practice

(e.g. an attempt to enter an adult height outside the range 1-3 meters is queried before it is added to the patient record). (103, 106)

Lastly, in order to ensure that no patient identifiable data is left in the record, the free-text comments field associated with each clinical event is filtered. (103) Only phrases which have been anonymised by EPIC are allowed to remain in the data that are released to researchers. These comments can be searched and anonymised by EPIC by hand if required for a particular study.

Other than the addition of these extra flagging fields and the exclusion of patient identifiable comments, there are no modifications made to the data. (103) The data are provided to researchers as full, coded demographic, medical and prescription details at individual patient level (table 3.1). Within these electronic records, clinical data such as diagnoses are stored as 5-byte Read-codes, drug prescriptions issued are stored using generic drug names, and clinical measurements and test results are linked to a coding system developed by InPS.

During the period covered by this study, practices that contributed to THIN were similar to national practices in terms of age (table 6.2), mortality and patient turnover (table 6.3), but tended to be larger (table 6.4), and practices from deprived areas may have been underrepresented. (86)

Table 6.1 THIN data supplied to researchers

1. Demographics (the PAT data table)

- Dates patients registered at practices
- Dates patients left practices
- Patient registration status (temporary/permanent)
- Year of birth
- Gender

THIN Data does not supply the following: name; exact address; exact date of birth; NHS number.

2. Diagnoses (the MED data table)

- All conditions and symptoms entered on the practice computer during consultations between the GP and patient. Medical conditions are recorded using the Read Clinical Classification version 2.
- Information on referrals to secondary care, including the specialty of the secondary care service.
- Secondary care information and other related information received by the practice may be entered retrospectively, including:
    - Details on hospital admissions
    - Discharge medication and diagnosis
    - Outpatient consultation diagnosis
    - Investigation and treatment outcomes

3. Prescribing (the THE data table)

- The GP typically issues prescriptions to patients using their computer: prescribing is logged into the system automatically. The prescribing recorded in the computer logs the drug prescribed using the Multilex coding system, which automatically creates therapy records for THIN.
- Acute treatments and medicines for a chronic condition can be temporally linked with a symptom or diagnosis although this is not comprehensive in THIN.
- Details of prescriptions from ongoing outpatient specialist care or over-the-counter drugs may be summarised by the GP, but the degree of information depends on its direct relevance to the patient.

4. Additional Health Information (the AHD data table)

- Commentary from the GP entered into free text fields. This can sometimes contain confidential or identifying information: THIN checks and ensures these comments have been anonymised before release to researchers.
- Information on lifestyle and health factors such as smoking and alcohol intake, where recorded by the practice.
- Tests and laboratory results are also accessible. Currently (2011) about 75% of THIN practices are electronically linked to pathology laboratories and receive test results electronically.

5. Socioeconomic data (the PVI data table)

- The majority of patients are linked to postcode-based socioeconomic, ethnicity and environmental indicators, for example, Townsend quintile.

Source**:** http://csdmruk.cegedim.com/our-data/data-content.shtml [accessed 6/11/2013]

Table 6.2 Comparison of age distribution of THIN practices with all practices in England, 2004

Age group	England*	THIN
0 – 4	5%	5%
5 – 14	12%	12%
15 – 44	43%	42%
45 – 64	24%	25%
65 – 74	8%	8%
75 – 84	6%	6%
85 +	2%	2%
Total	100%	100%

*Source: National data for England from NHS Executive, 2004.

Table 6.3 Comparison of practices contributing to THIN and national data: death rate and transfer out rate

	THIN	National
Crude death rate[1]	10.3/1000 patients	10.2/1000 persons
Proportion of patients transferring out each year	7.5% (171160/2276866)	7.7% (3.5m/45m) [2]

1 Region: UK. Source: Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. Informatics in Primary Care. 2004;12(3):171-7.
2 Region: England. Note: does not include patients who were removed from their practice list following death. Source: http://www.connectingforhealth.nhs.uk/about/case/npfitstatus.pdf (year not given).

Table 6.4 Comparison of THIN list size with QOF data for England, 2005

	Number of practices	Median list size	Interquartile range	
			Q1	Q3
THIN	315	7185	4514	10106
QOF*	8484	5396	3119	8259

*Source: National data from Quality and Outcomes Framework for England, 2004/05.

Table 6.5 Socioeconomic distribution of patients in THIN

Townsend quintile	Percentage of patients	As a percentage of patients with known socioeconomic status
1 (least deprived)	22	25
2	19	21
3	19	21
4	17	19
5 (most deprived)	12	13
Not known	11	-

Note: Townsend quintile is based on postcode of patient residence. Percentage of patients in each quintile would be 20% if THIN practices had same distribution as national practices.

6.3. Criteria used to identify eligible practices

Individual THIN practices were eligible for inclusion in this study from one year after the date the Vision practice software was installed to ensure that the practice was using the system to its full extent. (109, 132) They were also required to have used the system for an additional two years following this in order to provide a total minimum duration of continuous clinical data of at least three years from each practice. The increased risk of death or diabetes complications may take several years to emerge following diagnosis, so a minimum observation period of two years (following the diagnosis of diabetes for each case) from each practice was believed to be an appropriate requirement. (133)

Practices which were known to have gaps in their clinical data were also excluded. These issues were recorded by the database provider, EPIC, at the time of data collection (for example, the practice computer system was not functioning for a period of weeks, so patient care may not have been recorded electronically during this period). EPIC also supplied a date for each practice, before which each practice may not have routinely recorded patient deaths: this is known as the AMR (acceptable mortality recording) date. (132)

6.4. Criteria used to identify eligible cases

Patients were eligible for inclusion as incident cases of type 2 diabetes if they were registered in their practice for at least one year prior to the first Read-coded mention of diabetes, and if the diagnosis took place during the study period (1998-2003). The Read codes used in this definition are listed in appendix 6. The code lists and case definition were originally developed for use in earlier studies [section 4.6]. These were updated to include more recent Read codes after consultation with medically qualified colleagues prior to their use in this study. Potential cases were excluded if they met any of the additional criteria, listed below.

Exclusion criteria

1. Diabetes Read-code with a missing date in the patient record: it is possible that the diagnosis occurred at some unknown time prior to the first diabetes code with a valid date.
2. Patient's practice had their Vision software installed less than one year before the first mention of diabetes: the practice may not have been using their computer system to its full extent and there may be missing data on other variables of interest.
3. Women who were pregnant at the time of the diagnosis were also excluded, as they were likely to have gestational diabetes. Any later mention of diabetes in the same individuals, but not associated with a pregnancy, was included if it met the case definition.
4. Patients who were under the age of 35 at the time of diagnosis, or who were treated with insulin within one year of diagnosis were excluded, as they were likely to have type 1 diabetes. This was based on the advice of specialist diabetes colleagues.

5. Patients who died or left their practice within three months of diagnosis were excluded as their practice would not have had sufficient opportunity to begin long-term management of their diabetes, and were unlikely to have recorded clinical measurements recorded following diagnosis.

The results of this case identification process are presented in section 7.2.

6.5. Outcome definitions

An outcome was defined as the first occurrence of any of the following conditions. The number and proportion of cases with each of these outcomes at and following diabetes diagnosis is presented in section 7.3.

6.5.1 CHD

The definition of CHD included MI, angina and revascularisation surgery, and any Read code which specified CHD without mentioning any of these subtypes or a specific surgical procedure (appendix 6). The date of diagnosis for CHD was the date of the first occurrence of any of the above Read codes [section 4.6 and table 6.1].

6.5.2 Stroke

The definition of stroke included Read codes for ischaemic stroke, haemorrhagic stroke, and codes where stroke was specified, but subtype was not (appendix 6). It did not include transient ischaemic attack (TIA). The date of diagnosis for stroke was the date of the first occurrence of any of the above Read codes [section 4.6 and table 6.1].

6.5.3 CKD

Chronic kidney disease was identified using date of the earliest of the following three events: a single low eGFR recorded by the practice; Read-coded CKD (appendix 6); and Read-coded kidney dialysis (appendix 6) [section 4.6 and table 6.1]. The threshold for low eGFR was set at 60 ml/min/1.73m^2, following the 2005 UK CKD guidelines (table 6.6). Cases with an eGFR of less than 60 ml/min/1.73m^2 were therefore categorised as having Stage 3-5 kidney disease from the date of the first such record. The process for deriving eGFR from measured creatinine is described in the next section.

Table 6.6 Stages of chronic kidney disease

Stage	GFR range	Description
1	90+	Normal kidney function*
2	60-89	Mildly reduced kidney function*
3	30-59	Moderately reduced kidney function
4	15-29	Severely reduced kidney function
5	<15 or on dialysis	Very severe, or endstage kidney failure

Source: 2005 UK CKD Guidelines (http://www.renal.org/CKDguide/full/CKDprintedfullguide.pdf).

* These stages are only treated as CKD in the presence of other factors, e.g. proteinuria and were not included in the definition of CKD as a comorbidity or an outcome in this study.

6.5.4 Death

The date of death for patients whose practice registration status indicated that they had died was identified using the date of death field provided by THIN. This field was created by the database suppliers to give researchers a guide to the patient's date of death. (103) The date was derived using an algorithm which used data from several locations in patients' electronic record: the date of death recorded using a template for entering the fact and cause of death by practice staff; the date this template was filled out if no specific date of death was recorded; the date associated with a Read code which indicated that the patient had died; and lastly, if no other valid source can be found, the date the patient was recorded as having transferred out of their practice.

6.6. Baseline characteristics

6.6.1 Demographic

The age of each case at diagnosis of diabetes was estimated using their year of birth (year of diagnosis – year of birth) as full date of birth was not available to protect patient identity. Each patient was matched to a Townsend deprivation quintile (five groups, ranked from least deprived, quintile1, to most deprived, quintile 5) using their postcode of residence by THIN as part of the data collection process. (103) Some patients had multiple deprivation scores as they moved home while registered at their single practice. The deprivation quintile dated prior to their diabetes diagnosis was used as their baseline value in these instances. Broad geographical region was also collected by THIN as part of their data collection process. The exact location of practices was not made available in order to protect the identity of contributing practices. The last demographic factor of interest, patient sex, was recorded by practices at the time patients registered. The baseline demographic characteristics of eligible cases are presented in section 7.5.

6.6.2 Comorbidities

CHD, stroke and CKD at baseline were identified using the definitions described in the previous sections. As patients may have their creatinine estimated for the first time following diabetes diagnosis, restricting the definition to results recorded prior to this baseline date may

have resulted in an unacceptable level of missing data. Creatinine results up to three months after baseline were therefore included and used to estimate baseline CKD status. The frequency of each comorbidity at baseline among eligible cases is presented in section 7.5.

6.6.3 Smoking status

Each case was identified as a smoker or non-smoker on each date where their tobacco consumption or status was recorded, or when cessation advice or referral was offered. The Read and AHD codes used to identify smoking status are listed in appendix 6. The smoking status of each case at diabetes diagnosis (baseline) was initially identified using the last record of their smoking status before diagnosis, even if this was some years earlier. If the first record of their smoking status after diagnosis indicated that they were a smoker, then they were recategorised as a smoker at baseline. This was done even if their last recorded status before diagnosis indicated that they were a non-smoker as it was assumed that they did not begin smoking following their diagnosis. Cases with no smoking status recorded at any time were categorised as non-smokers at baseline. The baseline smoking status of eligible cases is presented in section 7.5, and the proportion who continued to smoke following diabetes diagnosis is presented in the appendix, in table A7.2.

The effect of this method was to group ex-smokers with non-smokers, even if they had ceased smoking the day before diabetes diagnosis. It might have been more congruent with classifications used in other risk models (e.g. Framingham) to have required ex-smokers to have quit for at least one year prior to diabetes diagnosis, but smoking status was not recorded at regular intervals in the healthy population in UK general practice during the study period.

(134) Patients recorded as being an ex-smoker in the year prior to diabetes diagnosis may have quit at any time between the preceding record of smoking status and that date: this would make it difficult to estimate their actual quit date accurately.

6.6.4 Identification of numeric values associated with clinical measurements

The clinical measurements of interest were eGFR, BMI, HbA_{1C}, total cholesterol and systolic blood pressure. These results were stored in the Additional Health Data (AHD) table in the THIN database (table 6.1). This table contained the date of the record, a Read code, an AHD code and a numeric value or values, along with value labels for each result (e.g. 'mmHg'). The Read- and AHD-codes for the events identified for each risk factor are listed in appendix 6. Any anonymised free text associated with one of the Read codes of interest was also searched for numeric results. Coded and free-text values were excluded if they lay outside a range of acceptable values. These ranges were set in consultation with clinical colleagues and are shown in table 6.7. The values associated with each label were also viewed as histograms as an additional data quality check (figures not shown). Results with labels other than those listed in table 6.7 were excluded if their distribution was not similar to results with these more standard value labels. The results of this search for additional clinical values in free-text are presented in section 7.4.

Table 6.7 Acceptable range for numeric values

Creatinine	25-1000 micromol/L (used in eGFR calculation)
Height	1-3m (if aged 18 years or over)
Weight	not required as BMI range applied after calculation
BMI	10-60 kg/m^2 (calculated from height and weight)
HbA_{1C}	2-20%
Total cholesterol	0.5-15 mmol/L
Systolic BP	50-300 mmHg

6.6.5 Calculation of eGFR

Kidney function was estimated using a formula based on measured creatinine and the age and sex of the case. (135) This derived value, estimated glomerular filtration rate (eGFR), is the recommended way to measure kidney function as creatinine level alone is affected by non-renal influences. (100) The eGFR equation corrects for some of these influences, and is more sensitive for the detection of CKD than serum creatinine and may be more accurate than creatinine clearance. (100) eGFR was created from recorded creatinine values for each case, using the abbreviated MDRD equation (135) :

eGFR = 186 x (creatinine / 88.4)-1.154 x (age)-0.203 x (0.742 if female) x (1.210 if black) ml/min/1.73m^2

All patients were assumed to be non-black, as individual data on ethnicity were not routinely available in electronic patient records from general practice at the time of the study. (136)

6.6.6 Calculation of BMI

BMI values were automatically generated within the Vision system each time the weight of a patient was recorded (provided a height value was available). (103) Although there are now some internal checks within the Vision system which highlight unexpectedly large or small values, at the time of data entry (103), it was observed that automatically generated BMIs included in the dataset provided by THIN still contained unfeasibly large values (e.g. BMI 141 kg/m^2), or were missing on occasion. Further visual inspection of the data showed that these values were based on incorrectly entered height or weight data (e.g. weight entered as 8684 kg instead of 86.84 kg), were based on height measurements taken when the patient was under 18 years, or missing if no height measurement was available at the time weight was measured. New BMI values were therefore calculated on each date where their weight was recorded. As height is not routinely measured at the time of each weight measurement, the last recorded height for each case was used if the measurement was taken when the case was aged over 18 years. (103)

6.6.7 Estimation of baseline values for clinical measurements

The level of each of the clinical measurements of interest was estimated at baseline for each case using a multilevel model which made use of all results recorded during their follow-up period. (137, 138) This was done because clinical risk factors such as cholesterol and blood pressure are a target for treatment once the initial diabetes diagnosis has been made, and

single measurements which are recorded in the months and years after the date of diagnosis may not, therefore, reflect their level at diagnosis.

In order to validate this complex modelling approach to estimating baseline values, the baseline estimates obtained from the multilevel models were compared with a simpler method: the mean of the values recorded immediately before diabetes diagnosis and during follow-up. For this comparison the cohort was restricted to cases who had the risk factor of interest recorded within 90 days of baseline to allow the value predicted for each case by the simpler and more complex model to be compared directly with the observed value recorded in their patient record. As it was not possible to generate an overall statistic to describe the proportion of variation in the data explained by a multilevel model (e.g. R^2), the residual sum of squares ($\Sigma(O-E)^2$: the observed minus the predicted value, squared, then summed) generated by each estimation method was compared using F-tests.

The results of the modelling of baseline clinical values and the validation are presented in section 7.4.

6.7. Development of risk prediction models

6.7.1 Introduction

This section presents the methods used to develop the four separate risk prediction models (CHD, stroke, CKD and death). The relationship between the baseline characteristics and outcomes was assessed using survival models as cases were followed up for differing intervals (until they left their practice, died or the end of the study period).

Each model excluded cases who developed the outcome of interest prior to, or in the first three months following diabetes diagnosis, and those who died or left their practice within three months of diagnosis. The remaining eligible cases were followed up from three months to a maximum of five years. Cases were censored at the earliest of the following dates: developed outcome of interest, left practice or died, last collection date from practice, and five years following diagnosis. Survival time, the time the case exited from the study was entered into each statistical model as the interval in years from the diagnosis of diabetes to the time they developed the outcome of interest, deregistered from their practice, or when five years had elapsed.

The results of this modelling are presented in sections 7.6 and 7.7.

6.7.2 Choice of survival model

The Weibull survival model was selected as it is frequently used to model survival data and, like other parametric models, makes more efficient use of data than a Cox model if its underlying assumption (that the fitted model follows a Weibull distribution) is met. (139) In particular, estimates of hazard ratios will be more precise than the equivalent Cox model if the survival data are observed to follow the Weibull, or other parametric distribution. (139) The Weibull distribution used in this study had two parameters, scale and shape, roughly equivalent to the intercept and slope in a linear regression. It is related to the simpler exponential distribution, which is also used to model survival, but can fit a wider range of situations. As with Cox and other survival models, the coefficient produced for each explanatory variable included in the Weibull model can be expressed as a hazard ratio (HR). The Weibull survival function, S(t), can be described as $S(t) = \exp(-\lambda t^{\gamma})$ where λ (lambda) is the scale parameter, γ (gamma) is the shape parameter and t is time. (139)

The underlying assumption, that the fitted model data followed a Weibull distribution, was assessed by comparing observed and expected failures using probability plots generated by the *pweibull* program within Stata. (140) The command fits a two-parameter Weibull model to the data and graphs the proportion of cases observed to fail at each point in time with the proportion predicted to fail from the model.

It was not possible to combine the expected failure times for each case on a single probability plot as they follow distributions determined by their individual pattern of covariates, such as age and sex. Therefore, to allow them to be compared as a unit with the observed failures,

they were back-transformed to a common distribution using their observed failure time, and their shape and scale parameters from the fitted Weibull model.

The results of this model checking are presented in section 7.8.

6.7.3 How predictors were included in the survival models

Continuous variables (eGFR, BMI, HbA_{1C}, total cholesterol and systolic blood pressure) were entered into the model in their original metric, centred on their mean value, if they met the proportional hazards assumption (see below). Binary and categorical variables were also checked to see if they met the proportional hazards assumption. Variables which did not meet this assumption were transformed, combined with other variables, or included in the models as a set of distinct covariates.

Comorbidities at baseline or during the first 3 months following diagnosis of diabetes were included as a series of binary covariates (i.e. one yes/no covariate per comorbidity). A gap of three months ensured that outcomes that occurred close to the diabetes diagnosis date, because the person was first assessed for the outcome just after they were diagnosed with diabetes (e.g. with CKD), were treated differently than events which may have been caused by diabetes itself. Each comorbidity was included as a separate predictor in each of the other outcome models. For example, stroke was included in the CHD outcome model as ischaemic stroke and CHD can share a common underlying pathology. (141, 142)

6.7.4 Checking of proportional hazards assumption

The assumption of proportional hazards for each of the predictor variables in the survival models was assessed visually using log-log graphs. The intention was that predictors that were not observed to be proportional over time would be mathematically transformed, entered into the model as a set of distinct variables, or combined in order to meet this assumption. A separate log-log graph was generated for each outcome (CHD, stroke, CKD and death) and predictor as survival may have differed from model to model for a particular outcome.

Log-log graphs compare the Kaplan-Meier (KM) estimate of the survivor function on the x-axis against survival time in logged form, hence the name of this type of graph. They present the same data that you would see in a standard KM graph (the lower the line is on the graph, the worse the survival probability), but the KM estimate and time are presented in this transformed metric to 'straighten out' the plots, allowing the viewer to more easily assess if the difference in survival for each level of each covariate was constant over time, i.e. that their hazards were proportional.

Hazards can be regarded as proportional if the line for each group remains roughly parallel with its neighbours over time, and their levels equally spaced. (139)In groups with large numbers of failures, the line tends to be smoother: in groups with smaller numbers of failures, for example in the youngest age groups, the line may be more erratic, and consequently it may be more difficult to make a judgment about proportionality.

The results of this checking are presented in section 7.6.2.

6.7.5 Handling of missing data

It is now widely accepted that complete case analyses and simple imputation of missing values are not appropriate methods for handling missing data especially where a significant proportion are missing. (50, 99, 143) Earlier studies with similar aims to the current study have either carried out complete case analyses, may have dropped incomplete variables from their analyses, or have carried forward previous or carried back later measurements for individuals. (55, 59, 144) The weakness of the complete case approaches is that they reduced the power of their study to explain their outcomes by dropping cases and variables which were associated with the outcomes of interest. They may also cause a systematic bias if the relationship between the dependent variables and the outcome for the cases that were retained in the analysis differed from the cases that were dropped.

Table 6.8 Types of missing data

Missing completely at random—There are no systematic differences between the missing values and the observed values. For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer.
Missing at random—Any systematic difference between the missing values and the observed values can be explained by differences in observed data. For example, missing blood pressure measurements may be lower than measured blood pressures but only because younger people may be more likely to have missing blood pressure measurements.
Missing not at random—Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values. For example, people with high blood pressure who are not adhering to treatment may be more likely to miss clinic appointments.

Source: Adapted from: BMJ. 2009 Jun 29;338:b2393. doi: 10.1136/bmj.b2393.

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR.

Statistical methods for handling missing data in health care databases have been in existence for some time (145), but were not available in widely used statistical software packages like Stata until recently. (143, 146) Multiple imputation of missing data (MI) has been used in other GP database studies, but is relatively complex and computationally intensive in large datasets: the high risk of inappropriately applying the process was demonstrated in one study published in the BMJ. (64, 147)

Although the use of multiple imputation may be preferable to simpler techniques, there are several stages of model building and decision making for each variable with missing data which may result in inaccurate estimates for the missing values being generated. (147, 148) Important assumptions have to be met if the imputed data are to be valid; most importantly that the missing data are 'missing at random' or 'missing completely at random' (table 6.8). This assumption is not possible to test directly, but knowledge of the reasons why the data are missing may support the use of multiple imputation, or suggest that it is not appropriate. (113)

Multiple imputation creates a number of imputed datasets which can contain different imputed values for each missing item. The individual imputed values are derived from one or more regression models within the imputation process: it is recommended that these should include all available explanatory variables to increase the plausibility of the missing at random assumption. (149) The main statistical analysis is then carried out on each of the imputed data sets in turn. The results of these analyses are then combined to produce a single set of results using 'Rubin's rules'. (150) In comparison with a single imputation process, multiple imputation generates larger standard errors that reflect the degree of uncertainty due to the use of imputation and better reflects the uncertainty due to missing values than a single imputed value. (151, 152)

Missing baseline values in this study were estimated using an imputation process known as multiple imputation by chained equations, implemented in Stata by Royston and colleagues. (146) This package (ICE) created a number of complete datasets, where missing baseline values were predicted using all available demographic and clinical data, and made an allowance for the imprecision of these predictions that was carried through to the final prediction models.

The results of this multiple imputation of missing baseline clinical values and deprivation quintile are presented in section 7.5.

6.8. External and internal validation of study results

This last section describes how the model results were validated. This consisted of comparisons with external studies and an overall statistical measure of each model's ability to explain variation in the outcomes observed in the study cohort.

6.8.1 Comparisons of results with other studies

Each of the demographic variables, clinical values and counts of outcomes used in this study were compared with the UKPDS RCT (1977-1991), and studies reporting on the Tayside diabetes register (1995-2004), the South Tees diabetes register (1994), and the Poole Diabetes Study (1996-1998) where data were available. (46, 88, 153-156) The eligibility criteria for the study cohort were adjusted to match these studies where possible, so that comparisons could be made on a like-for-like basis.

The hazard ratios for each model were also compared with the hazard ratios reported by other prediction models (tables 2.1 to 2.3). (44-46, 57-60) (67-81)

External comparisons with the results of other studies are discussed in section 8.4.

6.8.2 Overall proportion of variation explained by each model (R^2)

The proportion of variation explained by each model, also known as the coefficient of determination, was assessed using an implementation of the R^2 statistic called *str2ph*. (157) This was adapted from Nagelkerke's R^2 statistic for proportional hazard models for censored survival data. (158) The R^2 value can range from 0 to 1. A model with an R^2 of 1 would perfectly predict the outcome for each case.

These results are presented in section 7.7.

CHAPTER 7 RESULTS

7.1 Introduction

This chapter presents the results of this study. This includes the results of the case identification process and the baseline characteristics of the cases included in the study, and comparisons with other studies where possible. The results of the final prediction models themselves are then presented.

The first results presented are for the case identification process: these are presented in the form of a CONSORT chart (figure 7.1), The number and proportion of cases with each of the outcomes of interest (CHD, stroke, CKD and death) at baseline and in the first three months of follow-up (table 7.1), and in the period up to five years following the diagnosis of diabetes (table 7.2) are then presented. These data are also summarized in a single Kaplan-Meier type graph (figure 7.2).

The next section presents the results of the estimation and imputation of baseline clinical measurements of the study population. Some of the tables and figures can be found in the main appendix 7, to avoid presenting too many results in the text (tables A7.1, A7.4-A7.7; figures A7.1-A7.11). However, examples of each set of results are presented in the main text (table 7.3; figures 7.3-7.5). Table 7.5 then combines these results with the demographic, comorbidity and treatment data to summarise the baseline characteristics of the study population in a single table.

The final section presents the development of the prediction models and the model results themselves (tables 7.6 to 7.12). Two examples of the log-log plots used to check the proportion hazards assumption for each predictor in each model can be found in figures 7.5 and 7.6: the remainder can be found in the appendix (figures A7.4 to A7.11). The probability

plot used to check that the choice of a Weibull survival model was appropriate can be found in figure 7.7.

7.2 Cases identified

7.2.1 Overall results

Figure 7.1 presents the results of the case identification process. Overall, a total of 149492 potential cases were identified in the 300 general practices contributing to THIN at the time the study dataset was created. These practice patients had at least one of the selected Read codes for diabetes at some point in their record. A series of exclusions were then applied to these potential cases to ensure that the final set of data was as accurate and complete as possible (figure 7.1). Most of these exclusions (CONSORT items A-F) identify and exclude patients and practices (127601/149492 patients (85%)) that were never eligible for the study (the practices did not have research-quality data or the patients were diagnosed outside the study period). The remaining exclusion criteria (CONSORT item H) are more conventional (patients who did not have type 2 diabetes or who were not followed-up for a minimum period following diagnosis): these resulted in 1850/21891 patients (8.5%) being excluded from the study cohort.

The CONSORT chart (figure 7.1) describes the criteria in detail, but broadly potentially eligible patients were excluded if they or their practice had one of the features listed below.

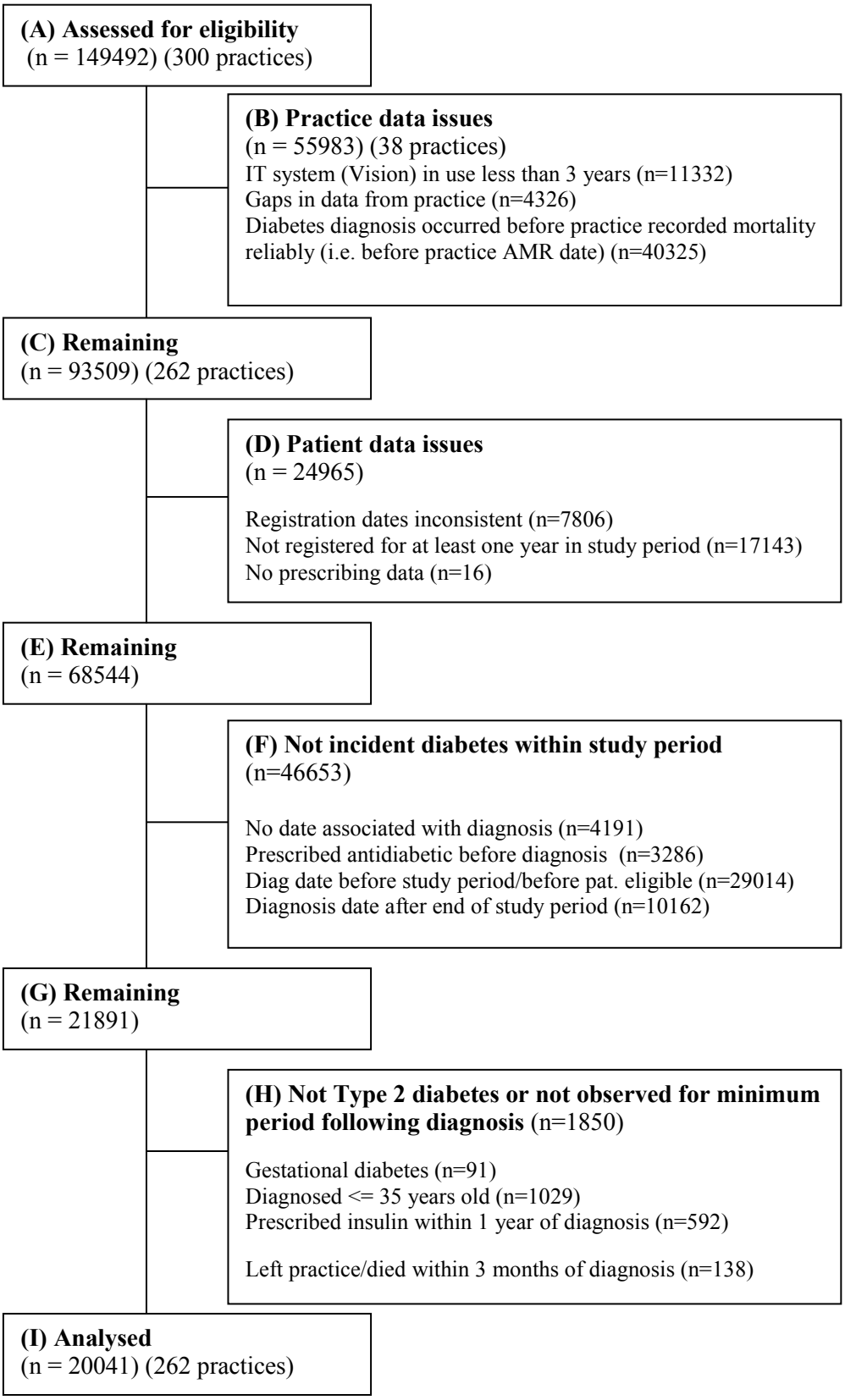
Data issues at practice and patient level [sections 6.2 and 6.3]

- Practice had insufficient experience of clinical computer system.
- Continuous use of system for minimum period.
- Patient diagnosed with diabetes before practice recording mortality reliably.
- Patient registration data inconsistent.
- Patient not registered for minimum period at practice
- No prescribing data for patient.

Eligibility issues [sections 6.2 and 6.3]

- Date of diabetes diagnosis not recorded.
- Patient prescribed antidiabetic treatment before diagnosis.
- Patient diagnosed before study period or less than one year after registration with practice.
- Patient diagnosed after end of study period.
- Patient likely to have type 1 diabetes.
- Patient left practice/died within 3 months of diabetes diagnosis.

Figure 7.1 Case identification: CONSORT chart



7.2.2 Exclusions due to practice data issues (CONSORT item B)

A total of 93509 patients remained after practices with specific data issues were excluded (37% excluded; 55983/149492). These issues were of two types: issues which affected the entire set of data from a practice (clinical computer systems in use less than three years and gaps in practice clinical data); and a single issue which depended on the apparent date of diagnosis of the practice patient (diabetes diagnosis before practice AMR date). The highest number of exclusions in this category was due to the latter issue: a total of 40325 cases were excluded to avoid under ascertainment of deaths.

7.2.3 Exclusions due to patient data issues (CONSORT item D)

The second group of exclusions (a further 17% of the original total; 24965/149492) centred on data issues which affected the records of individual patients, rather than the whole practice. A total of 24956 patients were excluded by these criteria, leaving 68544 patients (46% of the original total; 24956/149492). Almost all of those excluded were patients who had either inconsistent registration dates (a missing registration date, or a registration date which was later than their deregistration date), or were registered for less than one year during the study period. A small number of patients (n=16) were observed to have no prescriptions issued at any time in their record. They were excluded because they may have been training patients, set up by the practice to help staff to learn how to use the IT system.

7.2.4 Patients excluded because they were not incident diabetes within the study period (CONSORT item F)

The next group of patients to be excluded (31% of the original total; 46653/149492) were those who were diagnosed outside the study period, and those where a date of diagnosis could not be identified. Of the 46653 patients excluded at this point, the majority (84%; 39176/46653) were not diagnosed within the study period, and were therefore not incident cases of diabetes. The remainder (16%) either had no date associated with the first mention of diabetes) or were prescribed antidiabetic medication before the date of the first Read code for diabetes. As it can be several years before newly diagnosed cases of type 2 diabetes move from diet control to drug treatment, it was not possible to identify the true date of diagnosis of these patients.

7.2.5 Patients excluded because they had non-type 2 diabetes or insufficient follow-up time (CONSORT item H)

The last set of criteria applied excluded patients that were not cases of type 2 diabetes, or that were not followed up for a minimum period of time following diagnosis of type 2 diabetes (1%; 1850/149492). Of these: 91 had gestational diabetes; 1029 were age 35 years or less at the time of diagnosis (and therefore likely to be type 1 rather than type 2 diabetes); and 592 were prescribed insulin within one year of diagnosis (and also likely to have type 1 diabetes). At this point all remaining practice patients had incident type 2 diabetes, diagnosed within the

study period. The last criterion to be applied excluded 138 cases of type 2 diabetes that left their practice or died within three months of their diabetes diagnosis.

7.2.6 Summary of cases identified

Approximately 13% (20041/149492) of the potential cases assessed for eligibility were found to be eligible and therefore included in the analysis. Those that were excluded were excluded because it was not certain that the THIN database contained complete and contemporary records of their care at the time of diagnosis (data issues: 54% (80948/149492)), because they were diagnosed outside the study period (not incident diabetes: 31% (46653/149492)), or because they did not have type 2 diabetes mellitus or were not observed for a minimum of three months following diagnosis (1% (1850/149492)).

Excluding those who could never be eligible for inclusion in the study cohort (they were diagnosed outside study period or their practice did not have research-quality data) (127601/149492 patients), a total of 20041/21891 (91.5%) of patients with type 2 diabetes were eligible for inclusion in the study cohort.

7.3 Outcomes identified

7.3.1 Number of cases with CHD, stroke and CKD at diagnosis of diabetes

Approximately one-third of available cases (38%; 7561/20041) had one or more of the morbidity outcomes of interest before diabetes diagnosis (table 7.1). The most common comorbidity was CKD (22%), followed by CHD (20%) and stroke (6%).

7.3.2 Number of cases with CHD, stroke and CKD in first three months following diagnosis of diabetes

A further 1% (206/20041) of available cases were diagnosed with CHD in the 3 months following diabetes diagnosis. Stroke was a much less common outcome in the first three months, affecting less than half of 1% of cases. CKD was a relatively common outcome in this period, with approximately 1 in 20 (4.7%) new cases of diabetes being diagnosed with CKD. The 133 potential cases that died in the first three months following diabetes diagnosis were also not eligible for inclusion in the prediction models [section 7.2].

Table 7.1 Number of cases with CHD, stroke and CKD at baseline and in first three months following diabetes diagnosis

	Number of cases with outcome prior to diagnosis	(% of cases)
At baseline		
CHD	3969	(20)
Stroke	1240	(6)
CKD	4376	(22)
In first 3 months		
CHD	206	(1.0)
Stroke	75	(0.4)
CKD	950	(4.7)

Note: N=20041.

7.3.3 Number of cases eligible for each prediction model and number of outcomes observed between three months and five years following diagnosis of diabetes

All 20041 cases that survived and were still registered at their practice for at least three months after diagnosis were included in the survival model for death (table 7.2). The 3969 and the 206 cases who were known to have CHD in the period before or in the first three months following diabetes diagnosis were excluded from the survival model for CHD, leaving a total of 15861 cases eligible for inclusion. The 1240 cases and the 75 cases that had a stroke in the period before diabetes or in the first three months following diabetes were excluded from the survival model for stroke, leaving a total of 18726 cases eligible for

inclusion. The 4376 cases and the 950 cases who were known to have CKD in the period before diabetes or in the first three months following diabetes were excluded from the survival model for CKD, leaving a total of 14704 cases eligible for inclusion in the survival model for CKD.

Table 7.2 also shows that the CKD model had the greatest number of cases with an outcome of interest in the follow-up period, with 23% (3294/14704) of cases being diagnosed with CKD (stage 3+) in the 3 months to 5 years following diabetes diagnosis. The percentage of cases with the outcome of interest in the death, CHD and stroke models was 7.5%, 5.5% and 1.9% respectively.

Table 7.2 Number of cases eligible for inclusion in each prediction model and number of outcomes observed during follow-up

Model name	Number of cases included in model	Number of cases with outcome following diagnosis	(% of cases)	Mean age at diagnosis of diabetes	Years of follow-up Mean (SD)
CHD	15861	879	(5.5)	62	3.1 (1.3)
Stroke	18726	355	(1.9)	63	3.2 (1.3)
CKD	14704	3294	(22.4)	61	2.8 (1.4)
Death	20041	1502	(7.5)	64	3.2 (1.3)

Note: Cases were only eligible for inclusion in each model if they did not already have the outcome of interest at diabetes diagnosis or in the first three months following diagnosis.

7.3.4 Summary of outcomes observed at diagnosis of diabetes and in the following five years

Figure 7.2 shows the proportion of people in the full study cohort (N=20041) who were observed to have each of the study outcomes at diabetes diagnosis and the proportion who developed each outcome in the five years following diabetes diagnosis.

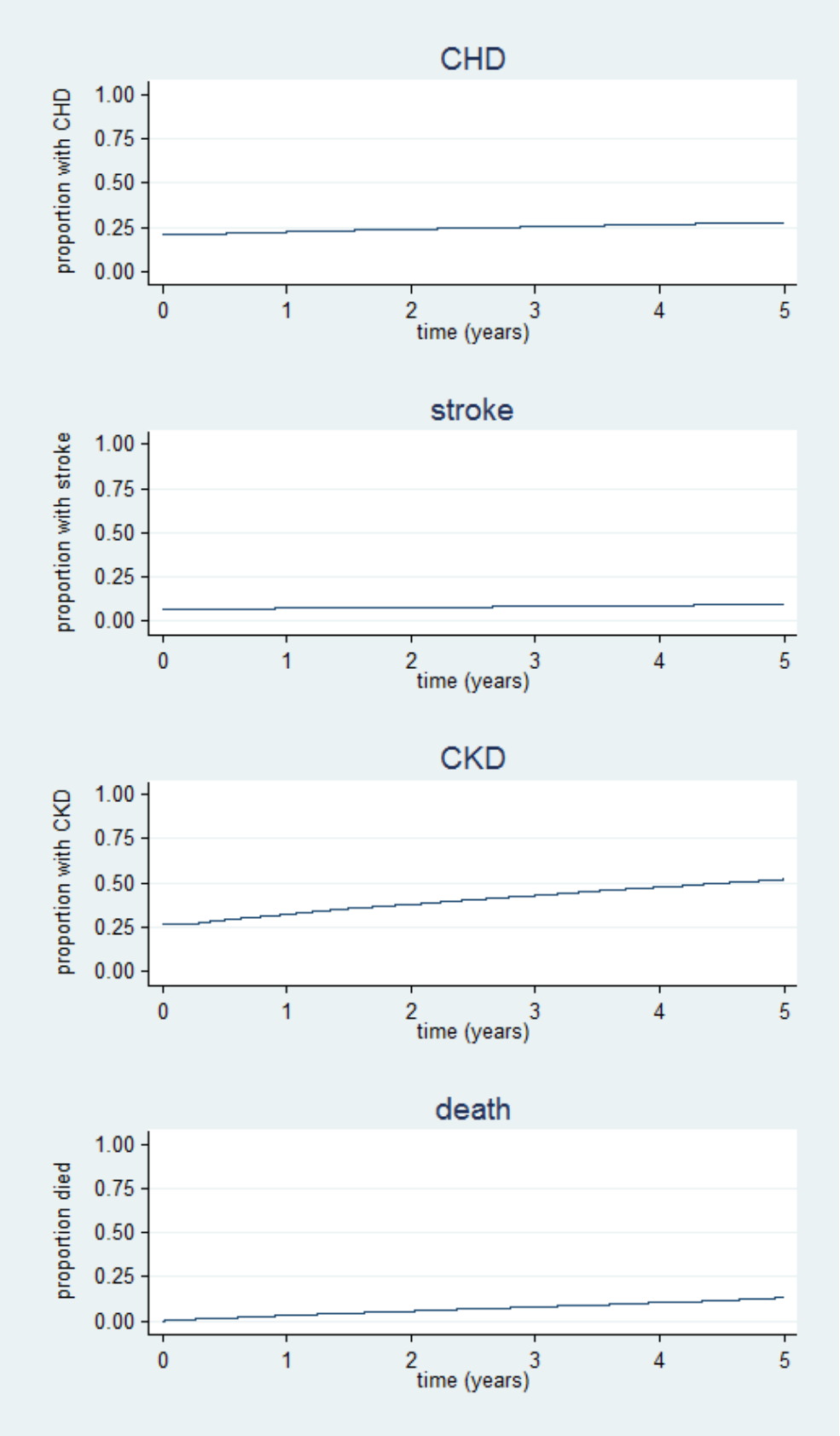
CHD: After adjusting for loss of cases to follow-up, 9% of the study cohort were estimated to have been diagnosed with, or died from, CHD in the five-year study period. Combining the proportion of cases with prior CHD from the original cohort of 20041 with the results of a Kaplan-Meier analysis, suggests that a total of 27% of cases with type 2 diabetes died from or were diagnosed with CHD by five years following diabetes diagnosis.

Stroke: A total of 355 of the 18726 cases included in the stroke outcome model were diagnosed with, or died from, stroke in the follow-up period. After adjusting for loss of cases to follow-up using the Kaplan-Meier failure function, 3% of this cohort were estimated to have been diagnosed with, or died from, stroke in the 5-year study period. The combined proportion of cases from the original cohort of 20041 who had experienced a fatal or non-fatal stroke by five years following diabetes diagnosis was therefore 9%.

CKD: A total of 3294 of the 14704 cases included in the CKD prediction model were diagnosed with CKD (stages 3-5) in the follow-up period. After adjusting for loss of cases to follow-up using the Kaplan-Meier failure function, 34% of this cohort were estimated to have had developed chronic kidney disease by the end of the 5-year study period. The proportion of cases from the original cohort of 20041 who had known CKD by five years following diabetes diagnosis was therefore 52%.

Death (all-cause): A total of 1502 of the 20041 cases included in the all-cause mortality prediction model died in the 5-year follow-up period. After adjusting for loss of cases to follow-up using the Kaplan-Meier failure function, 12% of the study cohort were estimated to have died in the five-year study period.

Figure 7.2 Proportion of study cohort with each outcome of interest at diabetes diagnosis and in the following five years



7.4 Estimation and imputation of baseline clinical measurements

The clinical measurements/risk factors of interest in this study were eGFR, BMI, HbA_{1C}, total cholesterol, and systolic BP. Few additional results were identified by searching for values entered into the patient record as free text (0.02% of HbA_{1C}s; 0.5% of total cholesterols and 1.7% of systolic BPs). The mean number of values recorded per case for each of the clinical measurements in the period between 30 days prior to diagnosis of diabetes and 5 years after diagnosis was as follows: HbA_{1C}=4.9; BMI(weight)=5.3; total cholesterol=4.1; systolic BP=9.3; eGFR(creatinine)=4.5. The baseline values of these clinical measurements were estimated using a multilevel model. The multilevel model for HbA_{1C} is shown in table 7.3. The remaining models (systolic BP, BMI, total cholesterol and eGFR) are shown in the appendix, in tables A7.4 to A7.7.

A graphical representation of the observed and modelled trajectory (using the multilevel models) for HbA_{1C} and systolic BP over time is presented in Figures 7.3 and 7.4. The equivalent figures for the other clinical values of interest (BMI, total cholesterol and eGFR) are presented in figures A7.1 to A7.3. These trajectories for HbA_{1C} and systolic BP are similar to those published as part of the UKPDS outcomes model (figure 7.5). For both studies and both clinical measurements, the modelled data for cases with the highest and lowest baseline values followed a funnel shaped path over the follow-up period, with the greatest changes in values in the earlier years.

The method used to estimate baseline values was internally validated by comparison with an alternative, simpler approach: mean value over the follow-up period. The results table is presented in table A7.1, rather than here to avoid presenting too many tables in this chapter.

For each clinical measure, the residual sum of squares (RSS) using the more complex multilevel model was between 24% and 66% lower than the simpler mean of the observed values. The difference in the RSS was highly statistically significant for each clinical value ($p<0.0001$), indicating that the multilevel models estimated baseline clinical values better than the simpler mean value method.

The mean estimated value for each of the measurements of interest is presented in table 7.4. HbA_{1C}, BMI, total cholesterol and eGFR were missing in 4%-7% of cases: BP was missing least often, for just over 1% of cases (260/20041). No baseline value could be estimated using a multilevel model for these cases. Instead, a baseline value for these measurements was generated for each of these cases using multiple imputation. The mean imputed value for each measurement is also presented in this table. The imputed data were very similar to the observed data for each clinical value, except for HbA_{1C} and eGFR. These differences are addressed here rather than in the discussion as they relate to the internal validity of the imputation process.

The mean observed and imputed baseline values for HbA_{1C} were 8.3% and 7.8%, respectively. However, cases with a recorded HbA_{1C} were twice as likely to be prescribed a drug to control their diabetes in the first two months following diagnosis (28% versus 12%, respectively). This suggests that it was reasonable to find that baseline HbA_{1C} was higher than those whose HbA_{1C} was imputed, and that the differences were not a result of any underestimation by the imputation process itself.

The mean observed and imputed baseline values for eGFR were 71 and 79 ml/min/1.73m2, respectively. As with HbA_{1C}, this difference may be explained in part by the characteristics of the cases with missing values. Although the two groups had similar baseline blood pressure

(144 and 146 mm/Hg), cases with no creatinine result (the basis of the eGFR calculation: if the creatinine level was missing, eGFR could not be calculated) present in their clinical record were less likely to be prescribed two blood-pressure lowering drugs recommended for use in chronic kidney disease. Angiotensin-converting enzyme inhibitors (ACE) were prescribed twice as often (53% and 27%) to cases with creatinine recorded, and angiotensin receptor blockers (ARB) were prescribed 11 times as often (11% and 1%) to cases with creatinine recorded in the follow-up period than cases with missing creatinine. This use of drug treatments suggests that their baseline kidney function was poorer, and therefore that their eGFR could reasonably be expected to be higher in those who were not treated. This assumes that the GP was more likely to record an abnormal eGFR, or that the first blood test was carried out in secondary care at the time of diagnosis.

A summary of the combined values for all these clinical values is presented in a single table in the next section (table 7.5), along with other baseline characteristics of the eligible cases.

Table 7.3 Multilevel model used to estimate baseline HbA_{1C}

HbA_{1c} (%)		**coefficient**	**p**	**95% CI**	
year of diagnosis	1998	0.214	<0.001	0.126	0.303
(reference year = 2000)	1999	0.015	0.700	-0.062	0.092
(1997 and 2004 omitted due	2001	-0.140	<0.001	-0.202	-0.079
to collinearity)	2002	-0.265	<0.001	-0.323	-0.207
	2003	-0.293	<0.001	-0.349	-0.236
age at diagnosis	35-44	0.196	<0.001	0.124	0.267
(reference age group = 55-64)	45-54	0.119	<0.001	0.067	0.170
	65-74	-0.140	<0.001	-0.185	-0.096
	75-84	-0.221	<0.001	-0.275	-0.167
	85-94	-0.259	<0.001	-0.360	-0.157
	95+	-0.358	0.155	-0.852	0.135
male		-0.036	0.038	-0.071	-0.002
smoker		0.010	0.467	-0.017	0.037
Townsend quintile	(least deprived) 1	-0.107	<0.001	-0.146	-0.068
(reference quintile = 3)	2	-0.069	<0.001	-0.107	-0.030
	4	-0.003	0.893	-0.041	0.035
	(most deprived) 5	0.083	<0.001	0.042	0.123
region	north	-0.009	0.669	-0.049	0.032
(reference = middle)	south	0.066	0.001	0.026	0.107
comorbidities	prior chd	0.006	0.799	-0.038	0.049
	prior chd	-0.075	<0.001	-0.112	-0.038
	prior stroke	-0.028	0.392	-0.091	0.036
drug treatments	insulin	-0.737	<0.001	-0.816	-0.659
	sulphonylurea	-0.400	<0.001	-0.423	-0.377
	biguanide	-0.427	<0.001	-0.446	-0.408
	acarbose	-0.199	0.013	-0.356	-0.042
	meglitinide	0.129	0.020	0.021	0.238
	glitazone	-0.485	<0.001	-0.526	-0.443
	statin	0.059	<0.001	0.039	0.078
	other lipid lowering	0.017	0.602	-0.046	0.079
	antianginal(excl. CCB)	0.012	0.546	-0.027	0.052
	aspirin	-0.042	<0.001	-0.065	-0.020
	OTC aspirin	-0.001	0.993	-0.249	0.247
	other antiplatelet	-0.100	0.001	-0.161	-0.039
	angiotensin-II receptor antagonist	0.020	0.277	-0.016	0.057
	ACE inhibitor	-0.105	<0.001	-0.128	-0.083
	alphablocker	-0.108	<0.001	-0.148	-0.068
	calcium channel blocker	-0.035	0.017	-0.063	-0.006
	diuretic	0.038	0.006	0.011	0.066
slope(change in HbA_{1C} per day)		0.0003	<0.001	0.0003	0.0003
time(CDF)		3.328	<0.001	3.284	3.372
intercept		7.520	<0.001	7.450	7.591

Figure 7.3 Observed and modelled HbA_{1C} over study period

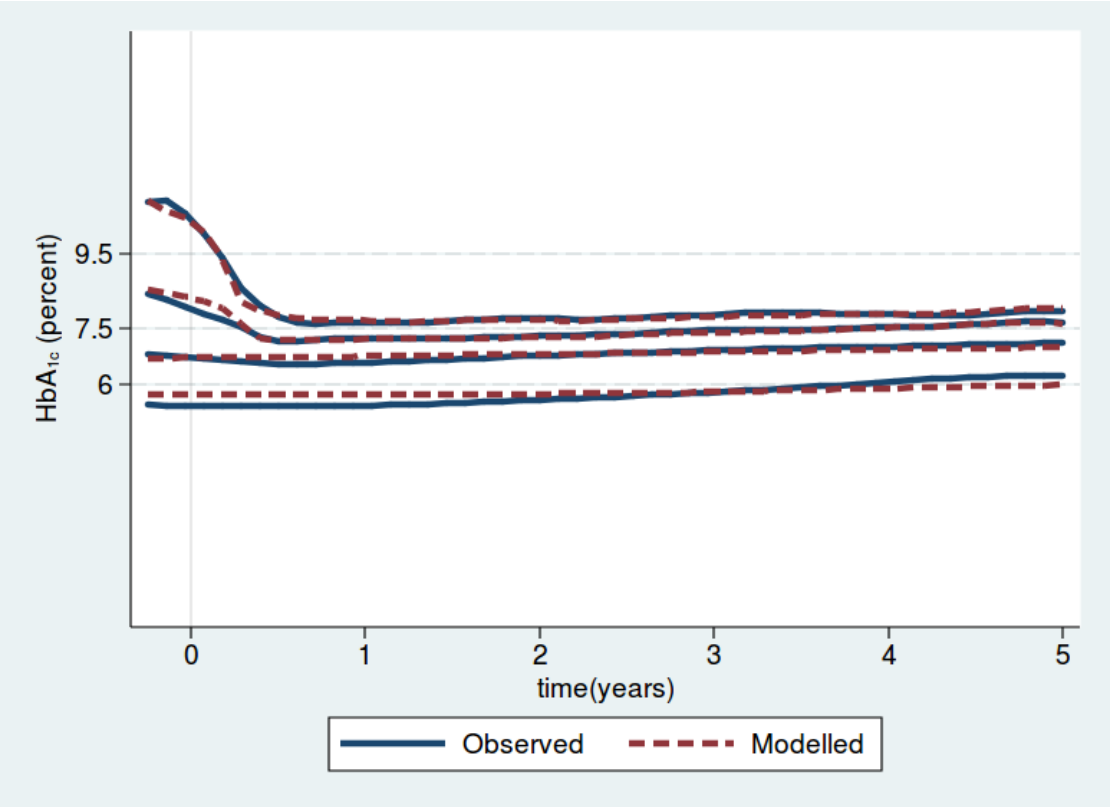


Figure 7.4 Observed and modelled systolic BP over study period

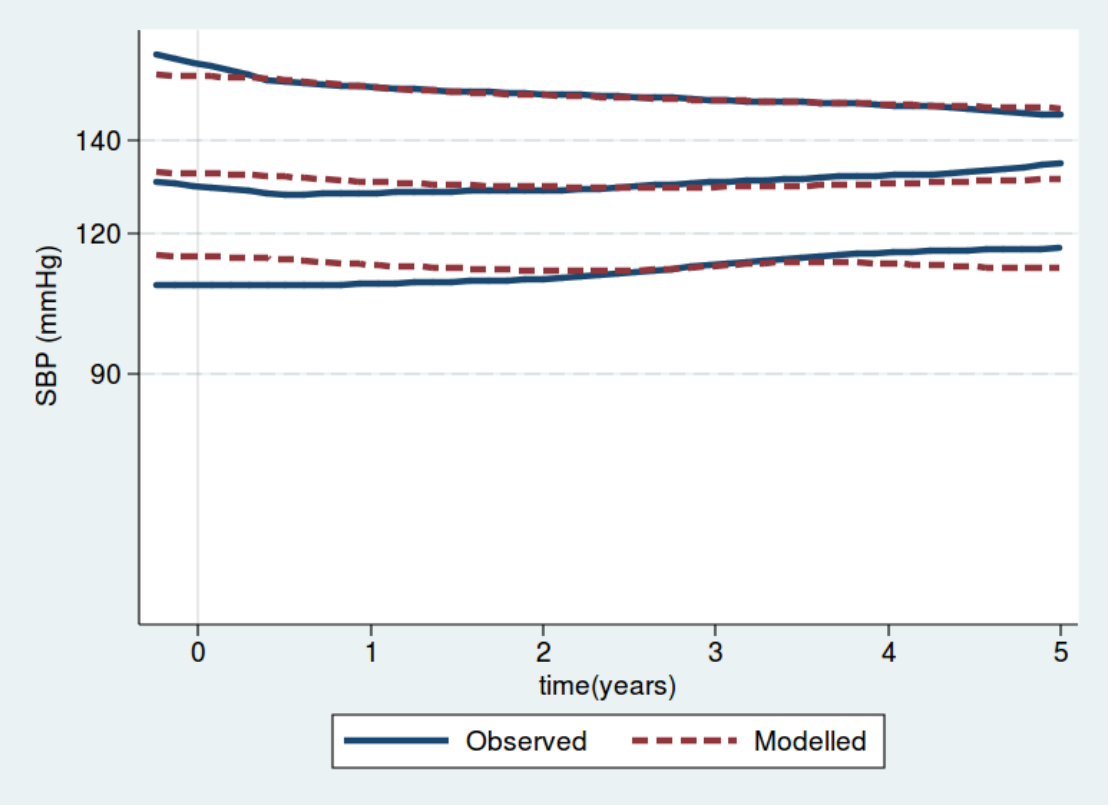
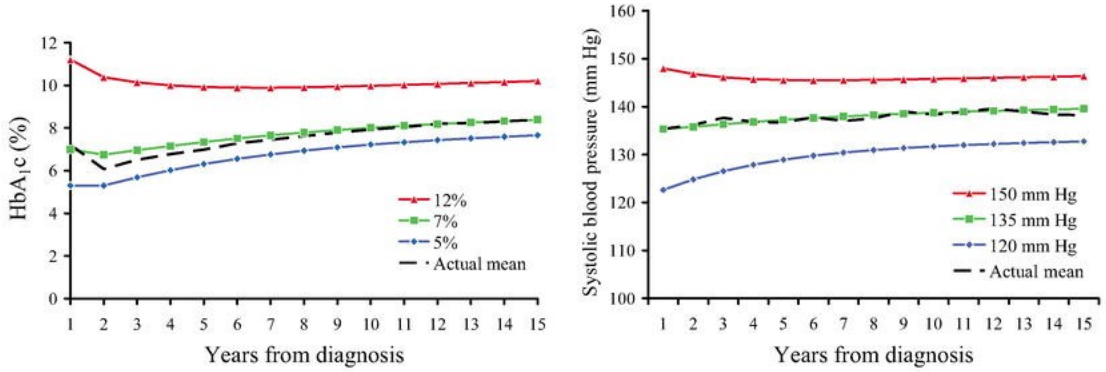


Figure 7.5 UKPDS observed and model estimated systolic blood pressure and HbA_{1C} over 15 years



Source: Clarke PM, Gray AM, Briggs A, Farmer AJ, Fenn P, Stevens RJ, Matthews DR, Stratton IM, Holman RR. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS 68) Diabetologia 2004;47:1747-1759.

Table 7.4 Distribution of clinical measurements: estimated using multilevel modelling, imputed using multiple imputation, and combined

		Cases	Mean value	(SD)
HbA_{1C}	estimated	19017	8.3	(1.9)
	imputed	1024	7.8	(1.3)
	combined	20041	8.2	(1.9)
BMI	estimated	18734	30.2	(5.8)
	imputed	1307	29.4	(4.3)
	combined	20041	30.1	(5.9)
Total cholesterol	estimated	19020	5.6	(0.9)
	imputed	1021	5.7	(0.6)
	combined	20041	5.6	(1.0)
SBP	estimated	19781	146	(13)
	imputed	260	145	(11)
	combined	20041	146	(13)
eGFR	estimated	19193	71	(15)
	imputed	848	80	(10)
	combined	20041	71	(15)

7.5 Baseline characteristics of eligible cases

Table 7.5 summarises the baseline demographic and clinical characteristics of the 20041 eligible cases included in this study in a single table for greater clarity. The demographic characteristics of this population and their baseline level of comorbidities were introduced earlier in this chapter, and are included here for completeness.

Townsend quintile was missing for missing for a total of 7% of cases and was imputed at the same time as the missing clinical values. Following imputation, cases were distributed relatively evenly across the deprivation quintiles, with the most deprived quintile slightly underrepresented (16% of cases). Smoking status at baseline could not be determined for less than 1% of cases: these had no information on smoking present at any point in their clinical record. Cases with missing smoking status were assumed to be non-smokers at baseline, rather than estimated using the multiple imputation process. A total of 24% of cases were identified as smokers at baseline using this method. Blood pressure lowering drugs were the most commonly prescribed group of cardiovascular drugs at diagnosis of diabetes (58%). Aspirin, prescribed by practices or bought over the counter by patients, was the next most commonly used drug (23% and 2%, respectively). Lipid-lowering drugs were the least frequent drug group, being prescribed to 14% of cases at baseline.

Table 7.5 Baseline characteristics of eligible cases

Cases		20041	
Male		10821	(54%)
Mean age in years		63.9	(SD 12.4)
Mean follow up time in years		3.2	(SD 1.5)
Deprivation	Q1 (least deprived)	21%	
	Q2	20%	
	Q3	23%	
	Q4	20%	
	Q5 (most deprived)	16%	
Comorbidities		**Cases**	**(%)**
	CHD	3969	(20)
	Stroke	1240	(6)
	CKD	4376	(22)
Current smokers		4890	(24%)
Clinical measurements		**Mean**	**(SD)**
	HbA_{1C} %	8.2	(1.9)
	BMI	30	(6)
	Total cholesterol	5.6	(1.0)
	Systolic BP	146	(13)
	eGFR	71	(15)
		Median	**(IQR)**
	HbA_{1C} %	7.7	(6.7 – 9.8)
	BMI	29	(26 - 33)
	Total cholesterol	5.6	(5.1 - 6.2)
	Systolic BP	145	(137 - 154)
	eGFR	72	(60 - 81)
Drug treatments		**Cases**	**(%)**
	BP lowering	11624	(58)
	Lipid lowering	2806	(14)
	Aspirin (prescribed)	4609	(23)
	Aspirin (OTC)	401	(2)

7.6 Development and checking of the statistical prediction models

7.6.1 Predictors included in the models

Individuals were entered into the model in their original metric where they met the proportional hazards assumption [section 6.6]. For example, age was entered as age at diagnosis of diabetes (centred on the mean age of cases), and gender was entered as one if male, zero if female.

Comorbidities at baseline or during the first 3 months following diagnosis of diabetes were included as a series of binary covariates (i.e. one yes/no covariate per comorbidity). A gap of 3 months ensured that outcomes that occurred close to the diabetes diagnosis date, because the person was first assessed for the outcome just after they were diagnosed with diabetes (e.g. with CKD), were treated differently than events which may have been caused by diabetes itself. The other outcome models (death, stroke) were treated in a similar way for simplicity. Each comorbidity was included as a separate covariate in each of the other outcome models. For example, stroke was included in the CHD outcome model as ischaemic stroke and CHD can share a common underlying pathology.

In the light of the observed sharp increases in the use of BP- and cholesterol-lowering drugs in the period immediately following diagnosis of diabetes (table A7.3) and the limited precision of estimated baseline SBP (limited to broad ranges of SBP, e.g. 90-119, 120-140, and 140+ mmHg) (figures A7.1 and A7.2), a conservative decision was made to enter baseline SBP and total cholesterol into the model as binary covariates. Specifically, cases with high SBP (140+ mmHg) (the point at which BP-lowering treatment is likely to begin) and

those who were already on BP lowering treatment at baseline were merged into a single group: cases who had past exposure to high blood pressure (as evidenced by their current treatment) and those who were currently exposed to high blood pressure. The comparison group, therefore, was cases who had normal SBP (<140 mmHg) and were not on BP-lowering treatment at baseline. A similar decision was made with respect to total cholesterol: cases with baseline total cholesterol of 4 mmol/L (the point at which cholesterol-lowering treatment is likely to begin) or higher were combined into a single group with cases who were on cholesterol-lowering treatment at baseline. The comparison group in this instance was cases who had a total cholesterol of less than 4 mmol/L and were not on cholesterol-lowering treatment at baseline.

Estimated GFR at baseline was not used in the all-cause mortality, CHD and stroke models as CKD was already entered as a binary comorbidity. Its effect on the overall R^2 (table 7.12) for the CKD model was, however, assessed in an additional CKD model where it was entered as a set of binary covariates, with cutoffs set to reflect the stages set out in the UK CKD guidelines (stage1: >=90 mL/min/1.73m2; stage 2: 60-89; stage 3: 30-59; and stages 4-5: <=29).

Lastly, baseline HbA_{1C} and BMI were entered as continuous covariates, centred on their mean value for the cohort. Unlike SBP and cholesterol, these were not likely to be treated at baseline, and the estimates of baseline values produced by the multilevel models appeared to tally well with the observed data (figures A7.2 and A7.3), allowing their effects to be assessed more conventionally.

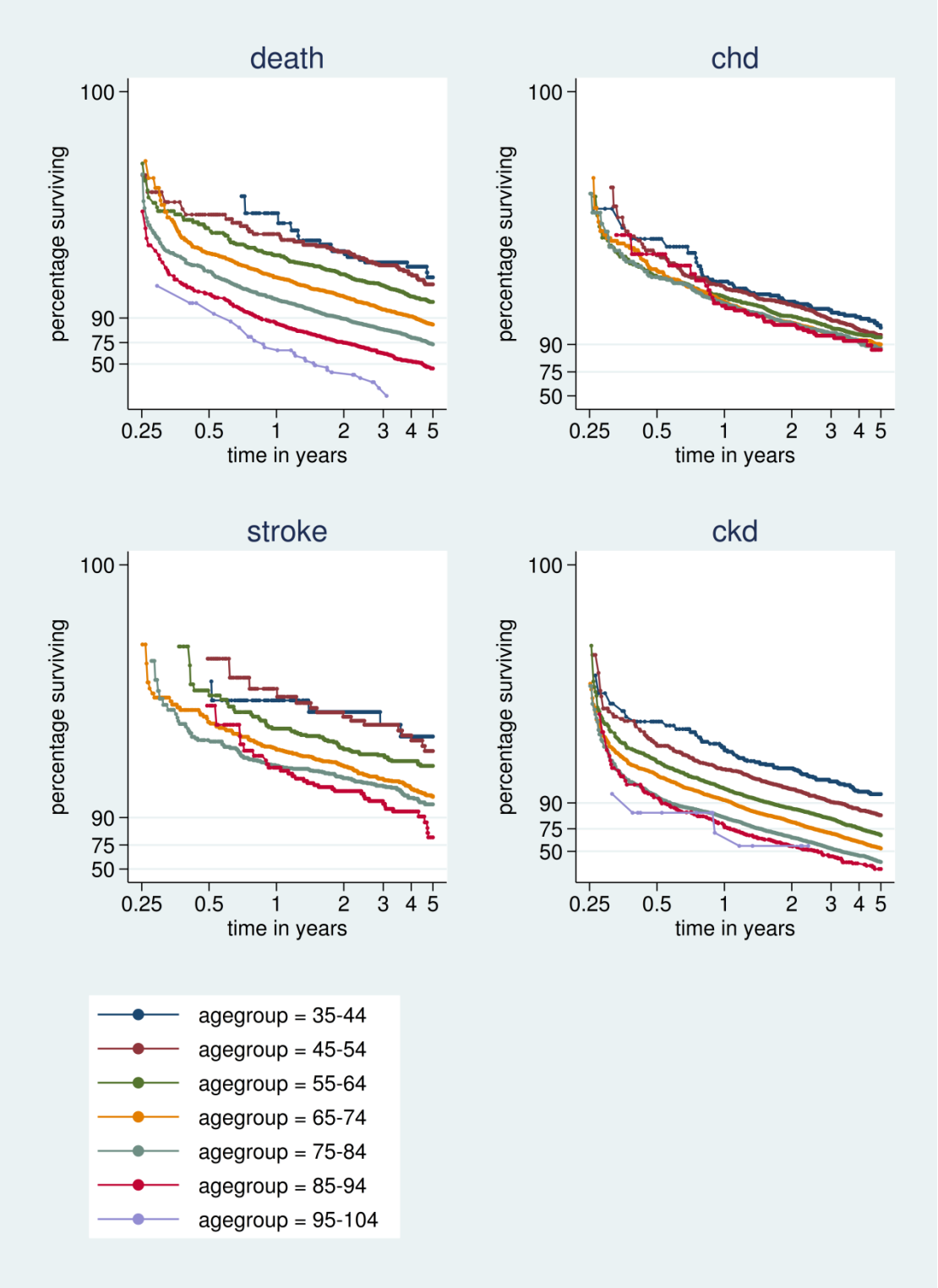
7.6.2 Model checking: proportional hazards assumption for each predictor

The purpose of testing the assumption of proportional hazards for each covariate included in each outcome model was to verify that they have been appropriately specified (parameterised) before being included in the survival models. The log-log plots for age and sex for each outcome of interest are presented in figures 7.6 and 7.7. The remaining plots are presented in the appendix, in figures A7.4 to A.7.11 to avoid presenting too many results in this chapter. Continuous covariates, such as age and BMI, were recast into distinct groups for this check, as the shape of each curve cannot readily be observed with a large number of closely spaced lines. Hazards can be regarded as proportional if the line for each group remains roughly parallel with its neighbours over time, and their levels equally spaced.

Age: Figure 7.6 shows a decreasing probability of survival with increasing age for each outcome model (death, CHD, stroke and CKD). The curves were parallel and there was a constant ratio between the curves for each age group over time, except for the youngest and oldest age groups, who had relatively small numbers of outcomes in the follow-up period. Even though it only appeared to meet the proportional hazards assumption for the intermediate age groups, age was kept as a continuous covariate in each survival model.

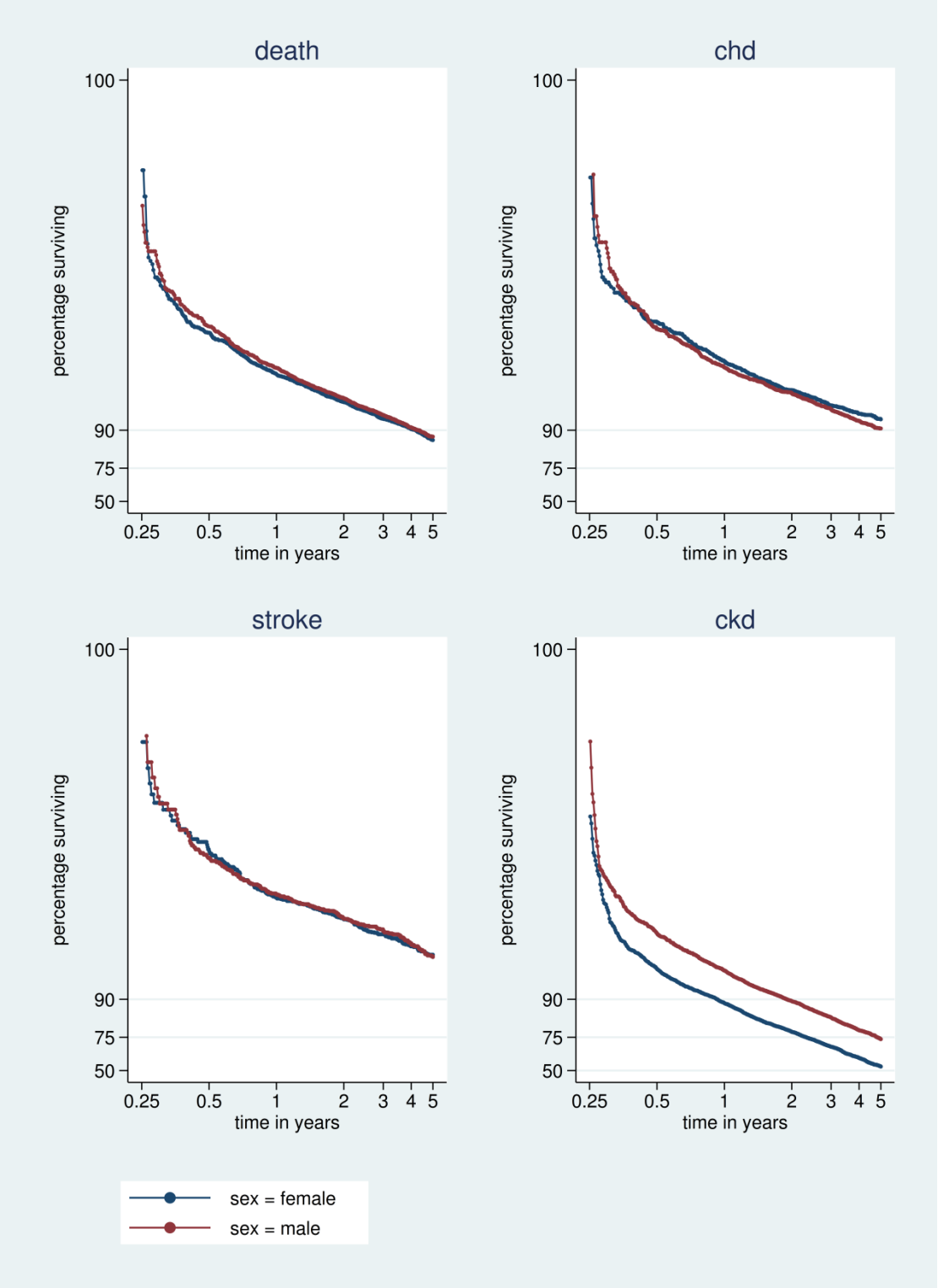
Sex: Figure 7.7 shows a similar probability of survival at each time point during follow-up for males and females for the death, CHD and stroke models. The survival probability differed in cases eligible for inclusion in the CKD model, but the groups remained parallel over time, suggesting that the hazard for each sex was proportional.

Figure 7.6 Log-log plots: age at diagnosis of diabetes



Note: Both axes are on logarithmic scales.

Figure 7.7 Log-log plots: sex



Note: Both axes are on logarithmic scales.

Smoking: Figure A7.4 shows that the risk of failure over time among cases who smoked at baseline was similar to that of non-smokers in each outcome. The proportional hazards assumption was met and smoking was included in each outcome model unaltered.

Comorbidities: Figures A7.5 to A7.7 show the plots for CHD, stroke and CKD as model predictors. The presence of each comorbidity was associated with a worse outcome for each of the outcomes of interest. There is no plot for a particular comorbidity (CHD, stroke or CKD) when it is the outcome of interest: these cases would not be eligible for a survival model where the outcome is the first ever diagnosis of CHD, stroke or CKD, respectively. Events in the period prior to diabetes diagnosis and in the first three months were combined as the study follow-up period began at three months following diagnosis. The plots for these comorbidities (CHD, stroke and CKD) remained roughly parallel over time. These figures show the proportion of cases observed to fail over time among those eligible for each outcome model. Each of these plots shows that the proportion surviving remains roughly parallel over the follow-up period. The proportional hazards assumption was, therefore, met and these covariates were entered into the survival models unaltered.

Clinical measurements and drugs: Figures A7.8 to A7.11 show the plots for the clinical value covariates, including those which were a product of the interaction with drug treatment (total cholesterol and systolic BP). The plots show that the covariates are roughly proportional over time and that the proportional hazards assumption was met. Drug treatments for high blood pressure and cholesterol at baseline, but not those initiated after diabetes diagnosis were included in each model for the following reasons:

- Levels of BP and lipid-lowering drug (lipid lowering) treatment were already relatively high at baseline (table 7.5) and treatment rates rapidly increased following diagnosis (table A7.3), reducing any difference in the exposure to harm in the follow-up period between

cases with a high but untreated baseline risk factor and those whose risk factor was low, or already treated.

- Any attempt to include treatments in the period following diagnosis, including those used to manage blood glucose, risked including an immortal time bias, as the case had to be alive to be prescribed a drug. (159)

7.6.3 Potential predictors considered but not included in the models

Year of diagnosis, geographical area and deprivation were also considered for inclusion in the survival models as they may plausibly predict the risk of death or diabetic complications.

It was decided not to include year of diagnosis in the survival analyses for the following reasons:

- Baseline HbA_{1C} was already included in the model, and the effect of lowering the threshold for diagnosis would be reflected in lower HbA_{1C} at diagnosis. The inclusion of year of diagnosis in addition to HbA_{1C} in survival models would, therefore, tend to lessen the apparent impact of HbA_{1C} in the risk of death and other outcomes. On balance, the effect of HbA_{1C} was of more interest than year of diagnosis.
- The study period for incident cases of diabetes was relatively short (1998-2003): the background age-specific risk of death, CHD, stroke and CKD in the wider population was unlikely to have changed in such a short period, lessening the need to control for its effect. (160)

- The number of practices contributing to the study cohort varied from year to year: any observed year-on-year differences in the risk of death and the other study outcomes may be due to differences between practices, rather than year of diabetes diagnosis.

On balance, as with year of diagnosis, a decision was made to drop geographical area and deprivation in favour of the clinically relevant variables for the reasons described below:

- Geographical region and deprivation are proxies for other, unmeasured, health related variables such as lifestyle, and access to and use of health services.
- They are both area-based measures, and do not refer to the individual case in the study cohort.
- Clinically relevant variables such as BMI and smoking refer directly to the individual case and are known to be on the causal pathway for the outcomes of interest in this study.

7.6.4 Interactions included in models

Baseline systolic BP and total cholesterol were categorised into binary covariates indicating if a case had a high level of each risk factor at baseline. High systolic BP and treatment with BP-lowering drugs were combined into a single binary covariate in order to meet the proportional hazards assumption, as they did not meet it individually. The rationale for this grouping, other than it meets the proportional hazards assumption, is that it creates two groups: those who were exposed to high BP in the past and the remaining cases who were not.

The effects of high baseline cholesterol and baseline lipid-lowering (lipid lowering) treatments were assessed independently as they both met the proportional hazards assumption.

As they were closely related, their combined effect was also assessed in an interaction term in each model.

7.7 Prediction model results

Tables 7.6 to 7.9 contain the results of the prediction model for each of the study outcomes: CHD, stroke, CKD and all-cause mortality. The first set of covariates in each table are demographic variables, the second are smoking/comorbidities, and the third are baseline clinical measurements and selected drug treatments. The hazard ratios (HRs), p-values, and the 95% confidence intervals for each of the HRs are displayed for each covariate included in each model. Table 7.11 combines these results into a single table to allow the HRs to be compared across models.

7.6.5 CHD prediction model results

Table 7.6 shows the results of the CHD model. Male sex (HR 1.36; 95% CI 1.18- 1.56) and higher age at diagnosis (HR 1.02 per year of age; 95% CI 1.01-1.03) were significantly associated with an increased risk of being diagnosed with CHD in the period from three months to five years following diabetes diagnosis. Smoking at diagnosis was also significant and increased the risk of CHD (HR 1.26; 95% CI 1.08-1.46). The effect of stroke and CKD, diagnosed at any time up to three months after diabetes had similar, but not statistically significant, HRs of 1.09 and 1.13, respectively (95% CI 0.84-1.42 and 0.97-1.33, respectively). Higher HbA_{1C}, higher BMI, and high SBP/BP treatment at baseline all appeared to increase the risk of CHD (HRs 1.07, 1.02 and 1.41, respectively) (95% CI 1.03-1.11, 1.00-1.03 and 1.13-1.76, respectively). High baseline cholesterol and lipid lowering treatment, however, were not statistically significant.

Table 7.6 CHD prediction model

		Hazard ratio	p	95% CI	
Male		1.36***	<0.001	1.18	1.56
Age at diagnosis of diabetes (per year)		1.02***	<0.001	1.01	1.03
Smoker at diagnosis		1.26**	0.002	1.08	1.46
Comorbidities prior to diabetes or in first 3 months following diabetes	Stroke	1.09	0.501	0.84	1.42
	CKD	1.13	0.117	0.97	1.33
Clinical measurements and treatments at diagnosis	HbA_{1C}[1]	1.07***	<0.001	1.03	1.11
	BMI[1]	1.02*	0.015	1.00	1.03
	SBP >= 140 mmHg or drug treated BP	1.41**	0.002	1.13	1.76
	Total cholesterol >= 4 mmol/L	1.38	0.530	0.50	3.83
	Total cholesterol >= 4 mmol/L and on lipid lowering drug	0.98	0.975	0.25	3.79
	On lipid lowering drug	1.79	0.396	0.47	6.84

* p<.05; ** p<.01; *** p<.001

1. Changes in hazard ratios are shown per 1% increase in HbA_{1C} and per 1 kg/m^2 increase in BMI.

Table 7.7 Stroke prediction model

		Hazard ratio	p	95% CI	
Male		1.15	0.190	0.93	1.42
Age at diagnosis of diabetes (per year)		1.06***	<0.001	1.05	1.07
Smoker at diagnosis		1.42**	0.005	1.11	1.81
Comorbidities prior to diabetes or in first 3 months following diabetes	CHD	1.50**	0.001	1.17	1.93
	CKD	1.17	0.152	0.94	1.48
Clinical measurements and treatments at diagnosis	HbA_{1C}[1]	1.01	0.756	0.95	1.07
	BMI[1]	0.99	0.235	0.96	1.01
	SBP >= 140 mmHg or drug treated BP	1.80*	0.015	1.22	2.90
	Total cholesterol >= 4 mmol/L	0.90	0.855	0.29	2.83
	Total cholesterol >= 4 mmol/L and on lipid lowering drug	1.22	0.781	0.29	5.08
	On lipid lowering drug	0.63	0.521	0.16	2.57

* p<.05; ** p<.01; *** p<.001

1. Changes in hazard ratios are shown per 1% increase in HbA_{1C} and per 1 kg/m^2 increase in BMI.

7.6.6 Stroke prediction model results

Increased age at diagnosis of diabetes was significantly associated with the risk of stroke in the first 5 years following this index event (HR 1.06; 95% CI 1.05-1.07), and male sex showed a positive hazard ratio, but was not statistically significant (HR 1.15; 95% CI 0.93-1.42) (table 7.7). Smokers at baseline also showed a significantly increased risk of stroke (HR 1.42; 95% CI 1.11-1.81). CHD, diagnosed at any time up to three months after baseline, showed a significant positive association with the risk of stroke (HR 1.50; 95% CI 1.17-1.93). Of the clinical measurements / treatments included in the model, only high SBP/BP treatment at baseline was significantly associated with the risk of stroke (HR 1.80; 95% CI 1.22-2.90).

7.6.7 CKD prediction model results

Unlike previous outcomes, the risk of CKD was lower in males than females (HR 0.52; 95% CI 0.48-0.56) (table 7.8). As seen in the other survival models, older age at diagnosis of diabetes appeared to be significantly associated with an increased of the outcome of interest (HR 1.06; 95% CI 1.06-1.06). Both CHD and stroke as comorbidities significantly increased the risk of CKD (HRs 1.21 and 1.14, respectively) (95% CI 1.11-1.33 for CHD), although the estimate for stroke came very close to non-significance (95% CI 1.001-1.289 for stroke). All the clinical measurement covariates with the exception of high total cholesterol and lipid lowering treatment at baseline were significantly associated with an increased risk of CKD. The HRs for HbA_{1C}, BMI, and high SBP/BP treatment at baseline were 1.03, 1.01 and 1.45, respectively (95% CI 1.01-1.05, 1.00-1.02 and 1.28-1.65, respectively).

Table 7.8 CKD prediction model

		Hazard ratio	p	95% CI	
Male		0.52***	<0.001	0.48	0.56
Age at diagnosis of diabetes (per year)		1.06***	<0.001	1.06	1.06
Smoker at diagnosis		1.10	0.019	1.02	1.19
Comorbidities prior to diabetes or in first 3 months following diabetes	CHD	1.21***	<0.001	1.11	1.33
	Stroke	1.14*	0.048	1.00	1.29
Clinical measurements and treatments at diagnosis	HbA_{1C}[1]	1.03**	0.002	1.01	1.05
	BMI[1]	1.01**	0.003	1.00	1.02
	SBP >= 140 mmHg or drug treated BP	1.45***	<0.001	1.28	1.65
	Total cholesterol >= 4 mmol/L	1.46	0.180	0.84	2.52
	Total cholesterol >= 4 mmol/L and on lipid lowering drug	0.64	0.164	0.34	1.20
	On lipid lowering drug	1.69	0.096	0.91	3.14

* p<.05; ** p<.01; *** p<.001

1. Changes in hazard ratios are shown per 1% increase in HbA_{1C} and per 1 kg/m^2 increase in BMI.

7.6.8 All-cause mortality prediction model results

Table 7.9 shows the results of the mortality model, that is, where the outcome of interest was all-cause mortality between 3 months and 5 year following a diagnosis of type 2 diabetes. Of the demographic covariates, male sex (HR 1.29), older age (HR 1.09), and smoking (HR 1.65) significantly increased the risk of death (95% CI 1.16-1.42, 1.09-1.10 and 1.46-1.87, respectively). All three comorbidities were significantly associated with an increased risk of death. In descending order of hazard ratio there were: CHD (HR 1.60), stroke (HR 1.47) and CKD (HR 1.33) (95% CI 1.40-1.80, 1.30-1.70 and 1.19-1.49, respectively). Of the clinical measurement at baseline only HbA_{1C}, BMI and lipid lowering treatment were significantly associated with the risk of death. Of these, only higher levels of HbA_{1C} were positively associated with increased risk of death (HR 1.09; 95% CI 1.06-1.12). Increased BMI appeared to be associated with a small reduction in risk per unit BMI (HR 0.98; 95% CI 0.97-0.99) in this model.

Table 7.10 shows the effect of total cholesterol and lipid-lowering treatment in more detail. Of these three related binary covariates (total cholesterol>=4 mmol/L, on lipid lowering drug, and the interaction term for these two covariates), only lipid lowering drug treatment was statistically significant. The relationship between total cholesterol and lipid lowering treatment and the risk of death was not simple or clear because of this: having either high cholesterol or being on lipid lowering treatment at baseline appeared to lower the risk of death. The effect of having both of these factors (treated but still high cholesterol), however was still associated with a lower risk of death than having normal and untreated cholesterol levels.

Table 7.9 All-cause mortality prediction model

		Hazard ratio	p	95% CI	
Male		1.29 ***	<0.001	1.16	1.42
Age at diagnosis of diabetes (per year)		1.09 ***	<0.001	1.09	1.10
Smoker at diagnosis		1.65 ***	<0.001	1.46	1.87
Comorbidities prior to diabetes or in first 3 months following diabetes	CHD	1.60 ***	<0.001	1.40	1.80
	Stroke	1.47 ***	<0.001	1.30	1.70
	CKD	1.33 ***	<0.001	1.19	1.49
Clinical measurements and treatments at diagnosis	HbA_{1C}[1]	1.09 ***	<0.001	1.06	1.12
	BMI[1]	0.98 **	0.002	0.97	0.99
	SBP >= 140 mmHg or drug treated BP	1.07	0.547	0.86	1.34
	Total cholesterol >= 4 mmol/L	0.61	0.080	0.35	1.06
	Total cholesterol >= 4 mmol/L and on lipid lowering drug	1.69	0.138	0.84	3.41
	On lipid lowering drug	0.41 *	0.012	0.21	0.82

* p<.05; ** p<.01; *** p<.001

1. Changes in hazard ratios are shown per 1% increase in HbA_{1C} and per 1 kg/m^2 increase in BMI.

Table 7.10 Combined effect of cholesterol and lipid lowering drugs on hazard ratios for all-cause mortality

		On lipid lowering treatment at diagnosis	
		no	yes
Total cholesterol at diagnosis	low (< 4 mmol/L)	1.00	0.41
	high (>= 4 mmol/L)	0.61	0.71

Note: Numbers are hazard ratios in comparison with cases with total cholesterol < 4 mmol/L and not on lipid lowering treatment at diagnosis of diabetes. This comparison group has a hazard ratio of 1.00.

7.6.9 All prediction model results combined

Table 7.11, below, presents the results of the models in a single table to facilitate later comparison of the results for each predictor across models.

Table 7.11 Hazard ratios for all prediction models

	CHD		Stroke		CKD		All-cause mortality	
Fixed risks								
Male	1.36	***	1.15		0.52	***	1.29	***
Age	1.02	***	1.06	***	1.06	***	1.10	***
CHD			1.50	**	1.21	***	1.58	***
Stroke	1.09				1.14	*	1.47	***
CKD	1.13		1.18				1.33	***
Modifiable risks								
Smoker	1.26	**	1.42	**	1.10	*	1.65	***
HbA_{1C}	1.07	***	1.01		1.03	**	1.09	***
BMI	1.02	*	0.99		1.01	**	0.98	**
High SBP / treated BP	1.41	**	1.80	*	1.45	***	1.07	
High total cholesterol	1.38		0.90		1.46		0.61	
Chol. high and treated	0.98		1.22		0.64		1.69	
Treated cholesterol	1.79		0.63		1.69		0.41	*

* p<.05; ** p<.01; *** p<.001
Note: Changes in hazard ratios are shown per 1% increase in HbA_{1C} and per 1 kg/m^2 increase in BMI.

7.7 Proportion of variation explained by each model

The proportion of variation in the data explained by each complete model was summarised using the R^2 statistic (table 7.12). Additional models were run for some outcomes in addition to the planned ones: these are shown in brackets.

The R^2 for each main outcome model varied between a maximum of 0.58 (all-cause mortality) and 0.09 (CHD). Both stroke and CKD performed similarly, with R^2 of 0.35 and 0.34, respectively.

Given the low R^2 for CHD, additional models were run to assess if this was due to the outcome including both hard and soft/intermediate outcomes. The R^2 for separate myocardial infarction (MI) and stable angina models performed better than the combined CHD model, with R^2 of 0.20 and 0.13, respectively.

The effect of including eGFR in the CKD outcome model was assessed separately from the main CKD model. The R^2 almost doubled from 0.34 to 0.66 after its inclusion.

Table 7.12 Proportion of variation in the data explained by the survival models

Model name	R^2
CHD	0.09
(MI only)	0.20
(angina only)	0.13
Stroke	0.35
CKD	0.34
(with eGFR)	0.66
All-cause mortality	0.58

7.8 Model checking: goodness of fit of Weibull model

The suitability of the choice of a Weibull distribution to model each of the study outcomes was assessed using probability plots (figure 7.8). These probability plots compared the proportion of observed failures (death, CHD, stroke, CKD) with those expected (fitted) using a separate Weibull model for each outcome of interest. The Weibull model was an appropriate choice of survival distribution if the plotted line (of observed versus fitted failures) lay along the diagonal, and did not grossly deviate from it at any point.

Death: There were 1502 deaths observed among the 20041 cases included in the death model. The plotted comparison of observed and expected deaths lay along the diagonal, showing a small deviation to either side of it at two points, before returning to the diagonal. Overall, the number of deaths derived from the model closely matched that observed in the data.

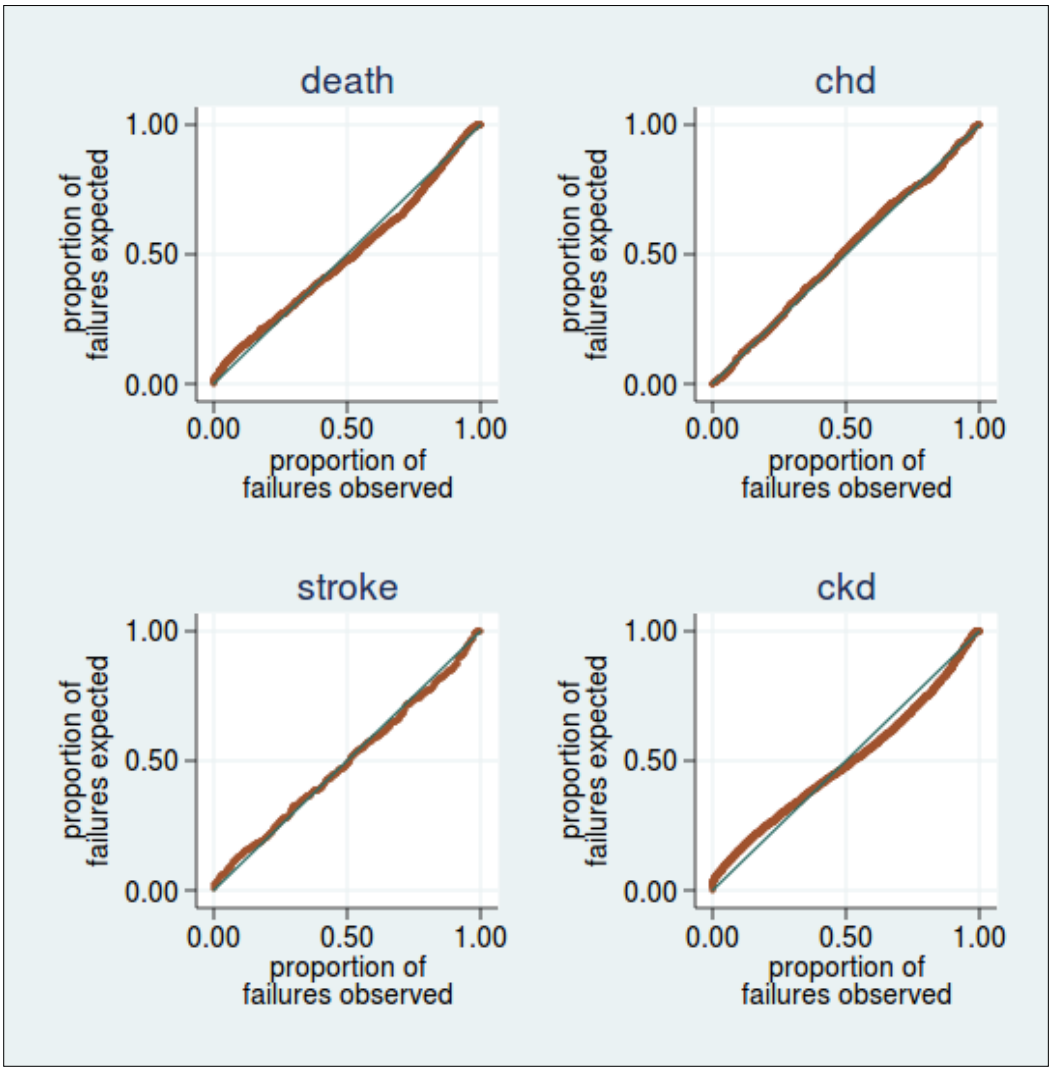
CHD: There were 879 cases of CHD diagnosed during follow-up among the 15861 people included in the CHD model. The plotted comparison of observed and expected occurrence of CHD lay along the diagonal, and did not systematically deviate from it at any point. Therefore, the number of expected CHD diagnoses derived from the model closely matched that observed in the data.

Stroke: There were 355 strokes observed during follow-up among the 18726 people included in the stroke model. As with the CHD model, the plotted comparison of observed and expected strokes lay along the diagonal, and did not systematically deviate from it at any point. Therefore, the number of expected strokes derived from the model closely matched that observed in the data.

CKD: There were 3294 cases of CKD observed during follow-up among the 14704 people included in the CKD model. The plotted comparison of observed and expected CKD occurrence lay along the diagonal, but did deviate from it at two points, before returning to the diagonal. At the point where 25% of the failures occurred in the data, the model predicted that approximately 30% of the cases would fail and at the point where 75% of failures were observed to occur, the model predicted about 70%. These differences are relatively small, and show the model both underestimating and overestimating failures at different points, but not grossly deviating from the centre line without returning to it. Therefore, the fitted Weibull model was an adequately close fit to the observed data, and accepted as a suitable choice for modelling CKD outcomes.

The Weibull model was accepted as a suitable choice of distribution to model each of the study outcomes. On the occasions where it deviated from the observed data, the differences were relatively small, and the model did not consistently over- or underestimate the number of outcomes observed.

Figure 7.8 Probability plots of observed failures vs. fitted Weibull model



CHAPTER 8 DISCUSSION

8.1 Introduction

This thesis concerns the development of four separate statistical models which can be used to predict the risk of coronary heart disease, stroke, chronic kidney disease, and all-cause mortality in the five years following a diagnosis of type 2 diabetes. The study used data from a large UK general practice database and included demographic variables, clinical predictors routinely recorded following diabetes diagnosis, and blood pressure and cholesterol-lowering treatments to populate the models.

This chapter discusses these models, focussing on their validity in comparison with existing models and their clinical utility. It begins with the main findings from each model [section 8.2] and the study's strengths and weaknesses [section 8.3]. This is followed by: detailed external comparisons to establish the generalisability of the study cohort in terms of demography and clinical features [section 8.4.1]; detailed comparisons of the estimates (hazard ratios) derived for each risk factor included in the current models with existing prediction models [section 8.4.2]; and comparisons of the population included and statistical methods with existing models [section 8.5], and approaches to missing data [section 8.6].

The last sections of this chapter discuss the implications of the results of this study. They identify which risk factors are most clinically important [section 8.7.1], and discuss the clinical utility of the models [section 8.7.2] and their implications for policy [section 8.8]. The very last section [section 8.9] describes the overall study conclusions.

8.2 Main findings

Age, sex and past medical history were important but fixed predictors of future risk. The hazard ratios for these non modifiable risk factors were: 1.02-1.10 for age (per year); 0.52-1.36 for male sex; 1.21-1.58 for CHD; 1.09-1.47 for stroke; and 1.13-1.33 for CKD. Key modifiable predictors were: smoking; weight; blood pressure; and glycaemic control. The hazard ratios for these risk factors were: current smoking 1.10-1.65; weight (per unit BMI) 0.98-1.02; blood pressure high or treated 1.07-1.80; and glycaemic control (HbA_{1C} %) 1.01-1.09. The proportion of variation explained by each model (R^2) was: CHD 0.09; stroke 0.35; CKD 0.34; and mortality 0.58.

The most clinically useful model might be the mortality model as it accounted for a large proportion of the variability in outcomes (R^2=0.58). This model found that age, sex and past medical history were associated with the risk of death, as were smoking, glycaemic control, BMI and high/treated blood pressure. The stroke and CKD models accounted for a moderate amount of the variation in outcomes observed (an R^2 of 0.35 and 0.34, respectively). The stroke model found that age, prior CHD, smoking and high/treated blood pressure were significant predictors of future stroke risk. The CKD model found that male gender, age, prior CHD and stroke were significant predictors of future CKD risk, as were smoking, glycaemic control, BMI and high/treated blood pressure. The CHD model had the smallest R^2 (0.09). Although it included known risk factors for CHD, the model accounted for little of the variation in outcomes between individuals and would not, therefore, be useful in clinical practice.

These results will be discussed in more detail in the individual sections below.

8.2.1 Prediction models

The main outputs from this thesis were the four prediction models. This section considers the results of each model in turn in terms of the individual risk factors that were significant in each model.

All-cause mortality prediction model: Perhaps the most successful of the models developed was that for all-cause mortality (table 7.11). This explained the majority of the variation (R^2= 0.58) in outcomes, more than any of the other models reported here (table 7.12). Possible explanations for this are discussed below [section 8.2.2]. This high R^2 also suggests that this model could be used in clinical practice to predict all-cause mortality risk, after appropriate validation [section 8.7.2].

People who smoked at the time of diagnosis had an increased risk of death of almost two-thirds (65%). Of the modifiable risk factors assessed across the four models, smoking was the only one to be significant in each case and had hazard ratios of sufficient size to make it perhaps the first target for intervention by clinicians.

There was a 9% increase in the risk of death in the follow-up period for every 1% increase in baseline HbA_{1C}. Higher HbA_{1C} at diagnosis of diabetes implies that the case had been exposed to high levels of blood glucose prior to diagnosis, possibly for several years, and therefore was at increased risk of vascular damage. (161) This may in turn increase the risk of death from CHD and stroke (and the risk of renal and eye disease, and neuropathy). An RCT based in eastern England which carried out population-based diabetes screening among high-risk individuals reported a lower baseline HbA_{1C} than was observed in this study (6.8% and

8.2%, respectively), suggesting that population screening might detect the disease at an earlier stage and so allow earlier intervention (table 7.5). (162) The key clinical message is that the association between higher baseline HbA_{1C} and serious outcomes observed in this study suggests a rationale for screening for diabetes, in that diagnosis earlier in the disease process could be associated with lower risk. NICE guidelines are already in place which address this. (163) These suggest that individuals at high risk of diabetes should have a blood test, and be reassessed every three years if they test negative. However a recent paper reporting 10-year outcomes from the ADDITION-Cambridge RCT found that a single round of screening did not reduce all-cause or cardiovascular mortality in the screened group. (164) Population screening for diabetes alone might not, therefore, be a cost-effective means to reduce these outcomes in the diabetic population, even if it led to a lower HbA_{1C} at diagnosis. As the authors suggested, a programme which assessed cardiovascular risk in addition screening for diabetes might be effective at reducing serious cardiovascular outcomes. This would benefit both the non-diabetic and diabetic population, and is a model used by NHS Health Check programme. (165)

The effect of increased BMI at baseline was also significant, but appeared to be protective: for every additional five units of BMI at baseline, the risk of death decreased by 10%. A similar protective effect of obesity on all-cause mortality was observed in a recent pooled analysis of five cohort studies. (166) The possible explanation for this unexpected effect given by the authors was that normal weight individuals with diabetes have a different genetic profile than overweight or obese individuals, and that these individuals are at risk of other diseases associated with higher mortality. (167) The observed effect of lower BMI on mortality in the current study could not be explained by three major diseases associated with higher mortality: overt CHD, stroke and CKD, as these were adjusted for in the model. If lower BMI

individuals are at higher risk of diseases other than diabetes it might, therefore, be due to disease other than these three (cancer, for example), or diseases which were diagnosed after diabetes. It may also be that newly diagnosed cases of type 2 diabetes with high BMI were younger, or diagnosed at an earlier stage in the disease than those with lower BMI, and therefore at decreased risk of death in the first five years. Higher BMI is a known risk factor for diabetes and this is reflected in the diabetes risk scores recommended by NICE to identify individuals who should be tested for diabetes. (163, 168-170) Therefore clinicians may consider diabetes more frequently in individuals with higher BMI. (163) Asymptomatic individuals may only have their diabetes diagnosed by random screening or when they are diagnosed with another disease. (163, 171-174)

Of the remaining clinical measurements / treatments included in the model, only high SBP/BP treatment at baseline showed the positive association that might be expected with the risk of death, although its effect was non-significant. This is not unexpected given that even in the BP Lowering Treatment Trialist' Collaborative meta-analysis of antihypertensive medication, the relationship between treatment and mortality (in comparison to major cardiovascular events) was not consistent. (175) The relationship between total cholesterol and lipid lowering treatment and the risk of death in the current study was not simple or clear: having either high cholesterol or being on lipid lowering treatment at baseline appeared to lower the risk of death (table 7.8). The effect of having both of these factors (treated and high cholesterol), however was still associated with a lower risk of death than being untreated and having a normal cholesterol level. Given that only one of these risk factors was statistically significant (the main effect for lipid lowering treatment), it is not clear if any valid inferences about their effect on death can be drawn from combining these results or whether these results were a chance effect or confounded by other factors not included in the model. If these hazard ratios

were an accurate reflection of reality, however, the explanation may be that patients with higher baseline cholesterol levels were more likely to be initiated on statins and, as a result, have a lower risk of future CVD, including stroke.

Of the fixed risk factors, males were at significantly higher risk of death in the follow-up period than females (29% more) (table 7.9). Prior comorbidity (CHD, stroke and CKD) each increased the risk of death by one-third to one-half. As the effect of these is additive, a person with all three of these comorbidities would have a 140% increased risk of death compared to a person free of all three comorbidities. Both of these findings are consistent with other data from the literature. (79-81, 176, 177)

CHD prediction model: In comparison to all cause mortality, the CHD model performed badly and most of the observed variability was not explained by the model (table 7.12). Of the modifiable risk factors, smoking was a prominent predictor of CHD risk: the risk of CHD for people who smoked at the time of diagnosis was 26% higher than for non-smokers, lower than the equivalent figure for all-cause mortality (65%) (table 7.11). As described in section 6.7.1, the CHD cohort excluded cases from the CHD survival analysis that had prior overt CHD (n=3969: approximately 20% of those included in the mortality cohort). This may have resulted in the exclusion of many cases with the highest levels of exposure to smoking, and therefore the highest risk of developing CHD: these cases may have developed CHD earlier than those with lower or no tobacco exposure. (178) Baseline data for smoking status was only presented for the mortality cohort in the results chapter (table 7.5), so this possible explanation cannot be supported here by evidence of a higher prevalence of smoking at baseline for the mortality cohort, compared with the CHD cohort.

Higher HbA_{1C}, higher BMI and high SBP/BP treatment at baseline were all associated with a significantly increased risk of CHD (a 7% and 2% increase per unit HbA_{1C} and BMI, respectively, and a 41% increase for high SBP/BP treatment). The effect of HbA_{1C} was similar to that for all-cause mortality (HR 1.09), and the effect of increased baseline BMI was positively associated with an increased risk of the outcome, unlike all-cause mortality, where increased BMI appeared to be protective. This suggests that HbA_{1C} and BMI should both be addressed in clinical practice in order to reduce CHD risk. Higher BMI also appeared to be a stronger predictor of CHD than stroke (HR 1.03 and 1.01, respectively) (table 7.11): this is consistent with evidence from the wider population that higher BMI has a greater effect on CHD risk than stroke risk. (179)

The effect of high SBP/BP treatment at baseline on risk of CHD (HR 1.41; 95% CI 1.13-1.76) was greater than that for mortality (HR 1.07; 95% CI 0.86-1.34), and unlike mortality, its effect was statistically significant. This suggests that BP is a more important target for treatment in order to reduce CHD risk than all-cause mortality, and fits well with trial data for the general population. (175)

The effect of high cholesterol and cholesterol-lowering drugs was to increase CHD risk, but the effect was non-significant and opposite to that observed for mortality. Again, this apparent difference may be explained in part by the exclusion of cases with overt CHD prior to, or within 3 months of diabetes diagnosis. Most CHD diagnoses occurred prior, or very close to, the diabetes diagnosis, rather than in the remaining follow-up period (3969 and 879, respectively). The effect of BP, cholesterol and drug treatment may have differed in those who developed overt CHD prior to the diagnosis of diabetes, compared to those who developed CHD following diabetes diagnosis, and this may have led to the observed differences between the CHD and mortality models for these risk factors. (180) Despite this

non-significant result for cholesterol on CHD risk, clinicians should not ignore this risk factor when attempting to reduce risk in people with Type 2 diabetes: there is ample evidence that cholesterol level is a risk factor from population-based trials. (41, 42)

In terms of non-modifiable risks, males were at significantly higher risk of CHD than females (36%), similar to that observed for all-cause mortality (table 7.11). This was also similar to the risk observed in the wider Framingham population (a 44% higher lifetime risk for males at age 70 years). (181) However the risk of CHD increased less with increased age at diagnosis than was observed in the all-cause mortality model: it increased by about 20% for each 10 year increase in age for CHD, but doubled for each 10 year increase for all-cause mortality. This may be due to the exclusion of cases with prior CHD from the CHD outcome model hence potentially resulting in a lower risk cohort. This highlights the difficulty in making comparisons across different models which are derived from different subsets of a larger cohort. It would have been possible to develop a set of models using a common cohort, allowing direct comparisons of hazard ratios. However, this would have resulted in the exclusion of at least 20% of patients (table 7.5), and produced models which could not be applied to a significant proportion of newly diagnosed patients with diabetes. On balance, it seemed better to ensure that the models were representative of the diabetic population than to ensure that hazard ratio estimates were directly comparable across models.

Overall the CHD model performed the worst of the three models. The explanation for the relatively low R^2 for this model was explored by running two additional sub models (table 7.12). The definition of CHD included both the soft/intermediate outcome of stable angina – which mainly relies on a clinical diagnosis – and the harder outcome of myocardial infarction (MI): the impact of each of the risk factors may therefore have differed for each of these two outcomes. The two sub models showed improved R^2 (0.20 and 0.13 for MI and stable angina,

respectively), suggesting that this was the case. Unfortunately, this approach could not be used in place of the main CHD model as only a proportion of the CHD outcomes could reliably be assigned to either MI or stable angina.

It is likely that the CHD-specific model presented here (R^2 0.09) could be improved if a reliable method could be found to split CHD outcomes into MI and angina. This might be achieved by linkage of GP and secondary care records or if there was an improvement in coding in general practice [section 8.3.2 weaknesses]. (112) Until such time as these can be achieved, this CHD model has relatively limited clinical utility [section 8.8 data issues].

Stroke prediction model: Considering modifiable risks, the effect of smoking at baseline was significantly associated with an increased risk of stroke, as it was with mortality and CHD (table 7.11). Smokers were at 42% higher risk of stroke in the period from 3 months to 5 years following the diagnosis of diabetes, reinforcing the importance of smoking cessation as a key clinical intervention.

Of the remaining modifiable risk factors included in the stroke model, only high SBP/BP treatment at baseline achieved statistical significance (95% CI 1.22-2.90). The effect of this was to increase the risk of stroke in the follow-up period by about 80%. This is almost twice as high as the HR observed in the CHD model (HRs 1.80 and 1.41, respectively). This is similar to the difference between the effect of blood pressure on CHD and stroke outcomes seen in studies on both non-diabetic and diabetic populations. (40, 182) The point estimate for the hazard ratio for blood pressure on future stroke was also the highest seen for any risk factor in any of the models and reinforces the importance of blood pressure control in preventing stroke.

The effect of high cholesterol, lipid lowering drugs and the interaction between the two were not significant, but, as with mortality, they showed a similarly counter-intuitive pattern of exposures that would be expected to increase risk, appearing to be protective. (41) It is not clear if this highlights a general issue inherent to the use of routine primary care patient data, or one specific to modelling chronic diseases like diabetes. It may be more likely to be the latter, as higher cholesterol levels predicted higher CVD risk in a study using one of the large GP databases in a statin-unexposed 'healthy' population (64, 109). In diabetes, patients with high cholesterol at diagnosis are likely to be treated with statins: the proportion prescribed lipid lowering drugs in this study rose from 19% at diagnosis to 42% at one year following diagnosis (table A7.3). High cholesterol at diabetes diagnosis, therefore, is likely to be associated with initiation of a statin in the period following diagnosis, and a subsequent reduction in risk. This may explain the apparent protective effect of high untreated cholesterol at baseline observed in this model (HR 0.90).

Unlike all the other models (table 7.11), the effect of male sex on stroke risk was not statistically significant (95% CI 0.93-1.42), though the direction of the hazard ratio was consistent with an increased risk of stroke for males (HR 1.15). The effect of age was significant, however, as it was with all the other models discussed so far: it increased by 6% for every additional year of age at diagnosis of diabetes.

Overall, the R^2 for the 5-year stroke model (R^2 0.35) was at a similar level to a 10-year CVD prediction model (QRISK1) developed for the wider population, which reported an R^2 of 0.33 and 0.36 for men and women respectively. (63) This suggests that this model could be used in clinical practice to predict stroke risk, after appropriate validation [section 8.7.2].

CKD prediction model: Higher HbA_{1C}, higher BMI and high SBP/BP treatment at baseline were all associated with a significant increased risk of CKD (a 3% and 1% increase per unit increase in HbA_{1C} and BMI, respectively, and a 45% increase in risk for high SBP/BP treatment). This pattern of risk is similar to that observed in the CHD outcome model (7%, 2% and 41% increases, respectively). Also, as observed in the CHD outcome model, the effect of high baseline cholesterol, lipid lowering treatment and their interaction was non-significant, but followed the expected direction: exposures which implied high levels of total cholesterol in the past were associated with an increased risk of CKD diagnosis in the follow-up period.

The effect of increased age at diagnosis was significant and increased the risk of being diagnosed with CKD in the follow-up period by 6% for each year of age (table 7.8). This is similar to that observed in the earlier outcome models (table 8.6), and may not be surprising given that the CKD risk is strongly related to age, even in the healthy population (71, 183). Unlike the other models, however, the effect of male gender was protective (and significant): males were at 48% lower risk of CKD in the follow-up period than females. This is consistent with the observed prevalence of stage 3-5 CKD in the general population, where it is approximately twice as common in females as males (7.3% and 3.5%, respectively). (184)

As seen in each of the other models (table 7.11), existing comorbidities, diagnosed at any time up to the first 3 months following diabetes diagnosis, increased the risk of the outcome (in this case CKD). CHD and stroke increased the risk of CKD by 21% and 14%, respectively.

Overall, the R^2 for the 5-year CKD model (R^2 0.34) was similar to that reported for a CVD risk model which was used in clinical practice. (63) This suggests that this model could be also be used in clinical practice to predict CKD risk, after appropriate validation [section

8.7.2] and perhaps after the inclusion of baseline eGFR as an additional predictor [section 8.2.2].

Comparison of hazard ratios between models in this study: The observed differences in hazard ratios for the same variables between the four models may have two explanations (table 7.11). Firstly, the cohorts are not directly comparable, as they each excluded cases with the outcome of interest. Also, as CHD, stroke and CKD can have a similar underlying disease process, atherosclerosis, the effect of excluding cases with an overt outcome of interest, may also exclude cases at risk of the other outcomes in this study. Secondly, other than differences caused by case selection, there are likely to be real differences in the effect of covariates, such as blood pressure levels at baseline, on each outcome. (40) Lastly, it is important to keep in mind when interpreting the hazard ratios for individual covariates, that it is the hazard ratio for that variable after all the other covariates have been taken into account. So, for example, the hazard ratio for smoking already takes into account the impact of sex and vice versa.

These models, however, do demonstrate that there are a set of known fixed and modifiable risk factors which predict future CVD and CKD risk, and risk of death, within 5 years following diabetes diagnosis. Age, gender and comorbid CHD, stroke and CKD predicted risk: where their hazard ratios were not statistically significant, they were in the expected direction. Smoking, higher HbA_{1C} and high/treated SBP were positively associated with increased risk of each outcome, though not always statistically significant. The effect of higher BMI predicted an increased risk of CHD and CKD outcomes, but was protective for stroke (but not statistically significant) and mortality. The clinical importance of each of these

risk factors is discussed below [section 8.7.1 what risk factors make most difference?; section 8.8 policy implications: clinical issues].

8.2.2 Differences in proportion of variation in outcomes explained by the models

The proportion of variation explained by each of the survival models (the R^2) varied widely, from a maximum of 0.58 for all-cause mortality, to a minimum of 0.09 for the CHD outcome model (table 7.12). For clinical populations like the study cohort, with wide age ranges and outcomes which are strongly age related, much of the explanatory power of the model will reside in the age and age-related variables (e.g. comorbidities present at diabetes diagnosis). Age at diagnosis was, therefore, probably the biggest contributor to the proportion of variation explained by each of the models.

A likely explanation for the relative success of the mortality model compared with the other outcomes, and CHD in particular, is the completeness and accuracy of recording of each outcome in primary care electronic patient records [section 4.6]. Morality in another UK primary care database was within 5% of national rates suggesting that the fact of death is well recorded in primary care. (130) The definition of CHD used in this thesis, however, was a composite outcome which included myocardial infarction (MI) and angina. A sensitivity analysis which developed separate models for these outcomes showed that the separate models predicted a much greater proportion of variation when separated (R^2: MI 0.20; angina 0.13) than when combined (R^2 0.09) (table 7.12), suggesting that the risk factors had a different effect on MI and angina. Even as separate models, they explained one-third or less of the variation in outcomes when compared with the mortality model. A recent paper on MI

which linked GP, secondary care and disease registry data found that 18% of the MIs recorded in the patients' primary care records could not be validated in the other data sources, and a similar proportion of MIs found in the secondary care and registry data could not be matched to events recorded in the primary care data. (112) This suggests that there is a substantial amount of underrecording and misrecording of MIs and, potentially, other acute events such as stroke in primary care electronic patient records. This is likely to have reduced the ability of the CHD and stroke models to accurately predict CHD and stroke outcomes, and led to the observed low R^2 observed for the CHD model. The CKD model explained a similar proportion of variation as the stroke model (stroke 0.35; CKD 0.34), but would have been substantially improved by the inclusion of eGFR at baseline (R^2=0.66) (table 7.12). This was greater than the R^2 for the mortality model, suggesting that any future versions of the model should include baseline eGFR as a predictor.

8.2.3 Clinical characteristics of people newly diagnosed with type 2 diabetes

A byproduct of the decision to develop risk models from the point of diagnosis of diabetes was that it also provided estimates of their clinical characteristics at this point (table 7.5). These results included comorbidities already present at diabetes diagnosis, clinical measurements and current drug treatments. As such it provides a snapshot of the characteristics of people diagnosed in the UK in the study period, 1998-2003, and can serve as a baseline for comparisons with more recent periods. As the outcomes of interest for the prediction models are mainly vascular-related, it is worth noting that one in five already had overt vascular-related disease at diabetes diagnosis (mainly CHD and CKD), and that some

drug treatments for primary or secondary prevention of CVD were already used by many practice patients by the time their diabetes was diagnosed (table 7.5). Of these the most common was BP lowering, with 6 in 10 cases already being treated at diagnosis. Patients already treated at baseline and those who began BP lowering treatment after diagnosis were likely to experience improved outcomes, as shown in previous studies in the diabetic and general population. (185, 186) Aspirin was also being used, but was less common at 1 in 4 cases, perhaps reflecting its use in secondary rather than primary prevention. Given the disagreement reported between systematic reviews of aspirin use in primary prevention in people with diabetes, it is uncertain if initiation for primary prevention prior to diabetes diagnosis would have any benefit over initiation after diagnosis. (24, 187) Lipid lowering treatments were the least commonly used CVD prevention drugs at baseline, used by just over 1 in 8 cases. Given that prolonged statin use is likely to produce larger absolute reductions in vascular events, patients initiated on a statin prior to diabetes diagnosis may have seen more reduction in outcomes than those who were initiated after diabetes diagnosis. (188)

8.3 Study strengths and weaknesses

8.3.1 Strengths

The study cohort was drawn from practices that were representative of the UK population and, therefore, was likely to be representative of patients with newly diagnosed diabetes. (189-192) Further, a minimal number of potential cases were excluded from the study cohort, increasing the likelihood that it was representative of the wider population of patients with type 2 diabetes. Only 8.5% of potentially eligible patients were excluded from the study cohort: where patients were excluded it was because they did not appear to have type 2 diabetes or were not followed-up for a minimum of three months following diabetes diagnosis.

The prediction models included risk factors which are known to predict the outcomes of interest and which are routinely recorded in general practice. This should ensure that the models are applicable to current UK general practice and that the data required to calculate these risks are available to practice staff. Some risk factors which may have predicted risk but were not routinely recorded in general practice during the period covered by this study were HDL cholesterol, waist:hip ratio and ethnicity. The lack of completeness in the recording of these risk factors prevented their inclusion in these models. The completeness of recording of HDL, and other laboratory results, will have improved since the introduction of electronic links with laboratories. (103) The recording of ethnicity for newly registered patients was incentivised in QOF, and has led to an improvement in the completeness of recording of ethnicity since the end of the period covered by this study. (193, 194) The recording of waist:

hip ratio in UK general practice for people with type 2 diabetes, however, is still not high: in a currently unpublished study using the THIN primary care database only 18% of patients with type 2 diabetes had a waist:hip ratio recorded at any time in their electronic patient record. The models presented here are, therefore, as complete as possible, although future updates, using more recent clinical data may allow the inclusion of HDL cholesterol and ethnicity as predictors.

The case definition was developed in earlier primary care database studies published by this author. (6, 7) The principal authors of that study had full access to the unanonymised free-text comments of practice staff. This allowed them to identify a set of Read codes with high sensitivity and specificity. This should have ensured that important groups of cases were not systematically excluded from the study cohort and have minimised the risk of including practice patients who did not have diabetes mellitus.

The prediction models developed in this study are likely to predict risk more accurately in patients with newly diagnosed (incident) type 2 diabetes than other models which were developed using prevalent diabetes cases. (44-46, 52-60, 76-81) This study showed that some risk factor levels, particularly HbA_{1C}, total cholesterol and systolic BP show steep declines in the first years after diagnosis, and that cases became more homogenous over time (figures 7.3, A7.2 and 7.4). Previous models which used risk factor data from the years following diagnosis are likely to systematically underestimate the level of these risk factors at diabetes diagnosis and will, therefore, overestimate the effect of these risk factors on outcomes [section 8.9.1 paragraph 4]. This implies that the models presented here could provide a more accurate estimate of risk in people with newly diagnosed diabetes than these other models, and would, therefore, be more useful in a clinical setting as they would help target treatment to those at highest risk.

8.3.2 Weaknesses

The time period covered by this study predates the introduction of the Quality and Outcomes Framework. Since the mid-2000s general practices in the UK have been financially incentivised to meet targets for the monitoring and treatment of patients with diabetes and other diseases which are associated with increased cardiovascular risk. This is reported to have improved the management of type 2 diabetes compared with the period before its introduction (although the initial rate of improvement was not sustained). (16) The clinical areas that were incentivised included the control of blood pressure, cholesterol, and HbA_{1C}. These three areas saw improvements in risk factor control from 1998 (pre-QOF) compared with 2005: a more than doubling in the proportion of patients meeting their target in the case of blood pressure and cholesterol. The prediction models developed in this study may therefore overestimate the underlying risk of outcomes in people diagnosed with diabetes in recent years: patients diagnosed in the post-QOF period will have achieved better control of risk factors, and therefore better outcomes following diagnosis than the patients who were included in this study.

The median follow-up period following diagnosis was relatively short at three years: this led to a decision to limit the prediction models to five years, to avoid making predictions past five years using data from relatively few patients. This may reduce the utility of the prediction models in clinical practice as many, though not all, clinical guidelines focus on the level of risk over 10 years. (195) Despite this, a five-year period for people in their 60s (the mean age at diabetes diagnosis) may be an appropriate time period over which to report results, and five year risk is used in the New Zealand risk guidelines. (195) The distinction between five and ten year risk may be less important than the distinction between short-term and lifetime risk,

and how risk is communicated to patients. (196-199) Short-term risk equations may also preferentially identify people who have already accrued substantial risk and miss others who might benefit from early preventative treatment, and risk estimates need to be understood by patients. (200-205) Identifying the most appropriate risk tools and methods for risk communication in patients with type 2 diabetes is beyond the scope of this thesis, but these papers suggest that the restriction of the risk models presented here to five years rather than ten is less important clinically than how and when the risks are presented, and how patients are engaged in their own care.

Patient turnover may have influenced the results reported in this thesis. In general practice turnover is around 7% per annum on average, though it can be as high as 25% in some areas. (206) Also, over half of all home moves in people aged 80 years or over between the 1991 and 2001 Censuses were from private to communal establishments and were often associated with the onset of a chronic illness. (207) Prediction models based on general practice data, where cases are censored when they leave their practice, rely on the assumption that the rate of failures (e.g. death) is the same in those who are censored as those who remain at their current practice until the end of the study period. The above national data suggests that this may not be a safe assumption, especially for the oldest age groups, and that the models developed in this study may have underestimated the risk of death and the association between baseline cardiovascular disease and death. Despite this potential weakness, however, the 5-year risk of death in the presence of baseline CVD was very high: 58% higher in patients with CHD than without, and 47% higher in patients with stroke than without (table 7.11). Work is currently in progress to link the THIN database to national death records on cause of death and hospital records, including events that occur after patient deregistration. It will, therefore, be possible in the future to ascertain if patients who are censored when they

deregister experience outcomes at a different rate than those who remain registered, and the extent of any systematic bias that this might have caused in prediction models developed using GP data.

Relatively few (18%, 3643/20041) of the cases in the study cohort had all their clinical measurements recorded (that is, coded in their electronic patient record) within 90 days of diabetes diagnosis [section 7.4]. The extent of this missing data meant that additional computationally intensive and time-consuming steps had to be taken in selecting appropriate methods for dealing with these missing values, increasing the time taken to develop the models and decreasing the precision of the estimates (HRs) reported [sections 6.6.3 to 6.6.7].

The prediction models presented in this study were developed in parallel for convenience. Separate models with their own sets of covariates and interactions may have better predicted the individual outcomes, but this would have increased the amount of time taken to build these models. As the main aim of the study was to demonstrate that risk factors routinely recorded in general practice predict risk in this clinical group, it was decided that a common set of predictors would be sufficient to achieve this. The CHD prediction model would also have been improved if it had been possible to separate all CHD outcomes into MI and angina subtypes (including those who had both subtypes), as would the CKD model if baseline eGFR had been included (table 7.12). Linked primary care and hospital inpatient and outpatient data has recently (2013) been made available to researchers using THIN. (107) These might allow researchers to distinguish between angina and MI in future models using this linked data source.

The direction of some hazard ratios in some of the prediction models has not been fully explained. Higher BMI appeared to be protective for all-cause mortality but increased both

CHD and CKD risk: further development of these risk models individually would allow non-linear effects of BMI to be tested (e.g. using fractional polynomials), though this might be a non-trivial task as missing BMI data was multiply imputed for some cases. (208) It may also be that this association is clinically plausible: this was discussed earlier in this chapter [section 8.2 all-cause mortality prediction model]. The effects of high baseline total cholesterol and cholesterol treatment were not statistically significant in three of the four models: the point estimates appeared to show increased CHD and CKD risk and a protective effect for stroke and all-cause mortality. Further development of these models separately, as with BMI, could include a non-linear effect of total cholesterol on each outcome if, like BMI, a way to include the multiply imputed missing data can be found. (208)

Although published after the literature searches carried out as part of this thesis, a recent BMJ paper on the validity of acute MI diagnoses recorded in primary care suggests that these diagnoses have a positive predictive value of only 92% when compared with a disease registry data. (112) Given its relevance to this thesis, it was decided to include it here as it has implications for the ascertainment of CHD as a comorbidity and outcome in studies using primary care data such as this. The authors conclude that linked primary care, death certification, hospital and disease registry data are required to avoid biased ascertainment of acute MI outcomes. These kinds of external data were not available at the time the models in this thesis were developed, but, as mentioned above, linked data is being introduced for use with THIN primary care records. Future studies could make use of some of these data sources to validate existing GP data or identify additional outcomes.

8.4 Comparison with other cohorts, models and study designs

This section compares the results of and methods used in the current study with previous studies with respect to the generalisability of its results to other UK patients with newly diagnosed type 2 diabetes and the external validity of the results.

The first set of comparisons are intended to establish the generalisability of the current study cohort to other patients with newly diagnosed Type 2 diabetes in terms of demography, comorbidity and frequency of outcomes. The studies used in these comparisons were UK-based where possible.

The next set of comparisons was intended to externally validate each prediction model, that is, the hazard ratios for the demographic and clinical predictors included in each model. The models used for comparison were based on cohorts of cases with newly diagnosed type 2 diabetes where possible.

The last part of this section addresses two major methodological differences between the current study and the other prediction model studies identified from the literature. The first of these differences was the decision to restrict the cohort to cases who were observed from the point of diabetes diagnosis. The second was the method for dealing with missing data adopted in the current study.

8.4.1 Cohort comparisons: baseline levels of risk factors and frequency of outcomes

In order to understand whether or not the results in this thesis are generalisable, baseline levels of risk factors and subsequent frequency of outcomes were compared to the literature. Tables 8.1 to 8.3 provide comparisons between the current study cohort and previous studies. These comparisons consist of the baseline demographic and clinical characteristics of people with newly diagnosed type 2 diabetes, and the frequency of selected outcomes at and following diabetes diagnosis. The present study cohort was restricted to specific subgroups where required to provide like-for-like comparisons between the published results and the other studies. Some external studies reported results for type 1 and type 2 diabetes combined, or data for prevalent rather than newly diagnosed diabetes cases. These are indicated in the tables, where applicable.

Demography - age and gender of cases: As can be seen from the hazard ratios in the models, age and sex are amongst the most important risk factors found. Table 8.1 compares the age and sex of the study cohort with results from external studies in order to judge generalisability and comparability of results. The mean age at diagnosis of diabetes in the current study cohort was the same as that for the Poole study (both were 64 years), and higher than that for the South Tees study (58 years). (153, 209) The younger age of the South Tees cohort is likely to be explained by the inclusion of cases with type 1 diabetes. The mean age at diagnosis in the current study was also similar or the same as that found in the UKPDS (55 and 53, respectively) and Tayside studies (63 and 62, respectively), once the current study cohort was restricted to match the external studies. (59, 133)

The proportion of cases that were male in the study cohort was similar to that found in the South Tees study (54% and 56%, respectively). (153) The proportion of cases that were male in the UKPDS and Tayside studies were also similar to that found in the current study, once the study cohort was restricted to match the external studies' inclusion criteria. (59, 133)

The similar results reported for this and other cohorts of patients with diabetes suggest that the study cohort was representative of the wider population of patients with type 2 diabetes in terms of age and gender. Given the limited age range included in the UKPDS, it is also likely that the current study cohort was in fact more representative of the wider population of people with type 2 diabetes. These imply that the models developed in this study would be useful in clinical practice: they would accurately estimate the risk at diabetes diagnosis of these outcomes in the wider population of people with type 2 diabetes.

Table 8.1 Comparisons with previous studies: age and gender of cases

	Study	Note	Study period		
				Mean	**(SD)**
Age	***Current study***	***(all ages)***	***1998-2003***	***64***	***(12)***
	South Tees (153)	*	1994	58	
	Poole (209)	(all ages)	1996-1998	64	(13)
	Current study	***(ages 35-65)***	***1998-2003***	***54.5***	***(8)***
	UKPDS (133)	(ages 25-65)	1977-1991	53.3	(9)
	Current study	***(excl. prior CHD)***	***1998-2003***	***62.6***	***(13)***
	Tayside (59)	(excl. prior CHD)	1995-2004	61.7	(12)
				Percentage of cases	
Male	***Current study***	***(all cases)***	***1998-2003***	***54***	
	South Tees (153)	*	1994	56	
	Current study	***(ages 35-65)***	***1998-2003***	***59***	
	UKPDS (133)	(ages 25-65)	1977-1991	61	
	Current study	***(excl. prior CHD)***	***1998-2003***	***52***	
	Tayside (59)	(excl. prior CHD)	1995-2004	52	

* Included type 1 diabetes. Included prevalent cases of diabetes.

Prevalence of comorbidities at diagnosis of diabetes: Co-morbidities had a large influence on prognosis in the models developed in this thesis, particularly in the mortality model. Table 8.2 compares the prevalence of CHD, stroke and CKD at diabetes diagnosis in the current study cohort with external studies.

The prevalence of CHD in the current study cohort (19.8%) was much higher than that reported for the DAI study in Italy (3.3%) and the study by Uusitupa in Finland (38%), but was very similar to the Ruigomez GPRD study, once the age range of the current study was restricted to match that used in the GPRD study (17.3% and 17.0%, respectively). (210-212)

The prevalence of stroke found in the French drug trial reported by Cathelineau was substantially lower than that found in the current study (1.6% and 6.2%, respectively). (213) After the restriction of the current study to the age range, the prevalence of stroke in the UKPDS was lower than that that found in the current study (0.8% and 3.0%, respectively). (46) However, it is not clear from the report if the 37 UKPDS stroke cases were the complete prevalent stroke population, or if others had been excluded at an earlier point. This may be likely as the original UKPDS study excluded cases with more than one previous major vascular event. (88) Lastly, the Ruigomez GPRD study reported a similar prevalence of prior stroke after restriction of the current cohort (4.7% and 4.4%, respectively). (212)

Table 8.2 Comparisons with previous studies: prevalence of comorbidities at diagnosis of diabetes

	Study	Note	Study period	Percentage of cases
CHD	***Current study***	***(all ages)***	***1998-2003***	***19.8***
	DAI, Italy (210)		1998-1999	3.3
	Uusitupa, Finland (211)		1979-1981	38
	Current study	***(ages 30-74)***	***1998-2003***	***17.0***
	Ruigomez, GPRD (212)	(ages 30-74)	1990-1992	17.3
Stroke	***Current study***	***(all ages)***	***1998-2003***	***6.2***
	Cathelineau, Fra. (213)		1992-1995	1.6
	Current study	***(ages 35-65)***	***1998-2003***	***3.0***
	UKPDS (46)	(ages 25-65)	1977-1991	0.8
	Current study	***(ages 30-74)***	***1998-2003***	***4.4***
	Ruigomez, GPRD (212)	(ages 30-74)	1990-1992	4.7
CKD	***Current study***	***(all ages)***	***1998-2003***	***22***
	South Tees (153)	*	1994	32
	Wolverhampton (214)	**	2002-2003	30
	Israel (215)	**	1999-2003	29
	Current study	***(ages 30-74)***	***1998-2003***	***16***
	Ruigomez, GPRD (212)	(ages 30-74)	1990-1992	1.3

* Included type 1 diabetes. Included prevalent cases of diabetes.
** Included prevalent cases of diabetes.

The prevalence of CKD at diagnosis of diabetes was 22% in the study cohort. The South Tees study reported a higher CKD prevalence (32%) in a cross-sectional cohort which included all diabetes cases in a region (both newly diagnosed (incident) diabetes cases and prevalent cases which had been diagnosed in previous years). (153) A similar CKD prevalence was reported in a study that used a regional diabetes register in Wolverhampton (30%), and among 269 prevalent diabetes cases in Israel (29%). (214, 215) These higher figures are consistent with the expected decrease in eGFR within individuals over time following a diagnosis of diabetes. The Ruigomez GPRD study reported a substantially lower prevalence of CKD than the current study, even after restriction of the current cohort (1.3% and 16%, respectively). (212) However, its definition of CKD required a diagnosis of albuminuria, proteinuria, renal failure, diabetic nephropathy or metabolic disorder, which will have only included cases where there was evidence of significant chronic kidney damage or established renal failure.

Although there were differences between the prevalence of comorbidities with non-UK studies, the prevalence of CHD, stroke and CKD was similar to that reported in more recent UK data. The strength of this evidence is weakened, however, for CHD and stroke, as both sets of results came from the GPRD: there was an overlap of approximately 50% in the practices contributing to that database and THIN.

Comorbidities were found to be highly predictive of each outcome in this study and were common at diabetes diagnosis. Unlike the clinical trials and observational studies from the UK and other countries, the models developed in this study included patients regardless of their health status and age, and provide UK-wide coverage. The current study models are more likely to be representative of patients seen in UK clinical practice in terms of comorbidity and age than these other studies, and more likely to accurately estimate baseline risk in the UK, resulting in models which can better predict risk in clinical practice.

Proportion of cases with outcomes in follow-up period: Table 8.3 compares the proportion of cases with each outcome in the current study with results reported for similar studies. A total of 5.5% of cases were diagnosed with or died from CHD during the follow up period in the current study (mean follow-up: 3.0 years). One study using the Tayside diabetes register reported that 5.3% of their incident cases of diabetes developed CHD during their 4.1 year follow-up period. (155) This is roughly equivalent to 3.6% over the period observed in the current study. This lower figure may be explained in part or in full by the younger age of the Tayside cohort (64 years and 55 years at diagnosis).

A total of 1.9% of cases were diagnosed or died from a stroke during the follow-up period (mean follow-up: 3 years). The estimated percentage of stroke outcomes in the UKPDS cohort was 1.2% over the first three years. (46) The equivalent percentage for the current study was close to this, once it was restricted to match the UKPDS age range (0.9%). The incidence of stroke also fell in the wider population between the period covered by the UKPDS and the current study (from 1981-84 to 2002-04). (216) This 29% decrease in incident stroke in the wider population is very similar to the difference between stroke occurrence in the UKPDS cohort and the current study (25%).

There was one non-UK study that reported changes in eGFR over time, but no studies reported the proportion of cases developing CKD after diagnosis of diabetes, so it was not possible to externally validate these particular results from the current study. (217)

A total of 7.5% of current study cases died during the follow-up period (mean follow-up: 3 years) (table 7.5). The Poole Diabetes Study reported that 20% of their incident cases of diabetes died during their 7.5 year follow-up period. (156) This is roughly equivalent to 8% of cases dying over a three year period, close to that observed in the current study.

The number of outcomes observed in the study cohort for CHD, stroke and all-cause mortality was similar to other published studies that used UK data: where differences were observed, they could be accounted for by differences in the age of each study population (CHD), or by differing study periods (stroke). This suggests that the study cohort was representative of the wider population of people with type 2 diabetes and, therefore, appropriate to use to assess prognosis.

Table 8.3 Comparisons with previous studies: proportion of cases with outcomes in follow-up period

	Study	Note	Study period	Percentage of cases
CHD	***Current study***	***(all ages)***	***1998-2003***	***5.5***
	Tayside (155)			3.6 *
Stroke	***Current study***	***(all ages)***	***1998-2003***	***1.9***
	Current study	(ages 35-65)	1998-2003	0.9
	UKPDS (46)	(ages 25-65)	1977-1991	1.2 **
CKD	(No external comparisons available)			
Death	***Current study***	***(all ages)***	***1998-2003***	***7.5***
	Poole (156)			8.0 ***

* Percentage estimated at 2.8 years based on percentage reported by study at 4.1 years (5.3%).
** Percentage estimated at 3 years based on percentage reported by study at 10.5 years (4.2%).
*** Percentage estimated at 3 years based on percentage reported by study at 7.5 years (20%).

8.4.2 Model comparisons: comparison of hazard ratios with other diabetes-specific and non-diabetes specific risk models

Introduction: This section attempts to externally validate the estimates of individual model predictors in the current study using comparisons with the estimates reported for earlier models. The hazard ratios for each of the current study models are listed in tables 8.4 to 8.7, along with equivalent results from other study models. The features of the comparison study populations are also listed where they differed from the current study in important ways, and a full list of the predictors included in each external model is provided as these also differed from study to study. It would have also been useful to compare the proportion of variation explained by each model (the R^2), but these data were only reported in the current study.

It should be noted that where a predictor reported by the external models does overlap with one from the current study, any dissimilarity in their values could be due to a number of factors. Each model is multivariate and includes a distinct set of predictors, spans different time periods, national populations and can be restricted to specific age-ranges. They also differ in terms of the target population: some are aimed at the wider population and include non-diabetics, and others include cases with type 1 diabetes or prevalent cases of type 2 diabetes. Therefore, unless those differences are repeated across a set of external models, it is unlikely that robust explanations for these differences can be identified. Despite this limitation, these external comparisons may provide general confirmatory or contradictory evidence about the expected importance or strength of specific predictors, and may also serve to highlight strengths or weaknesses of the current and external models in terms of generalisability to current clinical settings.

Three models derived from the UKPDS feature in the hazard ratio comparisons below. (45, 46, 81) They provide concrete examples of the difference between the current study cohort and external study cohorts. The UKPDS included cases diagnosed with diabetes between 1977 and 1991: the equivalent period for the current study was 1998-2003. The UKPDS was also restricted to cases aged 25 to 65 years at diabetes diagnosis: over half the incident cases of diabetes included in the current study cohort were aged over 65 years. The UKPDS models predict outcomes between 4 and up to 15 years following diabetes diagnosis: the current study was developed on outcomes observed between three months and five years following diagnosis.

In summary, the current study models for CHD, stroke and all-cause mortality included calendar periods, follow-up periods and age ranges not included in the UKPDS models: it is quite possible that estimates for an individual predictor may differ, yet both accurately described the effect of the predictor on the types of cases included in their respective cohort. It also indicates that the current study models for these outcomes are likely to be representative of the current UK population with type 2 diabetes, and, in particular, are valid for estimating risk in the years immediately following diabetes diagnosis, unlike the equivalent UKPDS risk models. It could, therefore, be argued that the models derived in this thesis are more pertinent to current clinical practice than other models including those from UKPDS.

CHD: Five CHD risk prediction models specific to people with diabetes were identified from the literature (table 8.4). (44, 45, 57, 59, 60) None of these models used the full set of covariates included in the current study model. All but one (45) included prevalent diabetes cases, and two included younger cases only (aged 45-64 and 25-65 years, respectively). (45, 60) Comparisons were, therefore, restricted to studies which reported hazard ratios for one or more overlapping predictors.

Three studies reported hazard ratios for age, male sex and current smoking. (45, 57, 59) One of these studies also reported a hazard ratio for HbA_{1C}. (45) A further one reported hazard ratios for current smoking and BMI. (60) The values reported by the Donnan model (59) for this and all common predictors were opposite to that reported by all other studies, including the current study. It may be that this is related to their reporting model coefficients rather than more readily interpretable estimates such as hazard ratios. No further comparisons with this study will be made in this text to avoid unnecessary repetition of its counterintuitive results.

The hazard ratio for male sex in the current study was 1.36 (95% CI 1.18-1.56): this was in the same direction but differed in level from the results reported for the two other studies: Yang (HR 2.01; 95% CI 1.66-2.63) and Stevens (HR 1.69; 95% CI 1.52-1.88). (45, 57) This was lower than the hazard ratio reported by both studies. However the Yang model was developed with a Hong Kong population which may not be directly comparable with the UK population and the 95% CI estimated by the current study overlapped with that of the model reported by Stevens suggesting that the observed differences were not statistically significant.

The hazard ratio for each additional year of age in the current study was 1.02: this was similar to the results reported by both studies that reported comparable data. (45, 57) The Yang model (57) reported a hazard ratio of 1.03 (95% CI 1.01-1.04) for age and the Stevens model reported a hazard ratio of 1.06 (95% CI 1.05-1.07). The estimate for the current study did not lie within the 95% confidence interval reported by the Stevens model. The inclusion of stroke as a predictor in the current model, but not in the Stevens model may account for some of this difference: the risk of stroke increases with age, and it's inclusion in a multivariate model with age would tend to reduce the hazard ratio for age itself.

The hazard ratio for smokers in the current study was 1.26. The hazard ratio for the other models also indicated that smoking increased the risk of CHD (45, 57, 60). The estimate reported by Yang was 1.55 (96% CI 1.08-2.22), by Folsom was 1.05 for males and 1.57 for females (95%CIs nor reported); and by Stevens was 1.35 (95% CI 1.11-1.59), respectively. In each case the estimate reported by the current study lies within the 95% confidence interval reported by the comparison studies or the gender-specific estimates that they reported.

Comparable results for HbA_{1C} and BMI were reported by a single study each. (45, 60) The hazard ratio for a 1% increase in HbA_{1C} in the current study was 1.07. This is close to, but still outside, the lower boundary of the 95% confidence interval for the estimate reported by Stevens (HR 1.18; 95% CI 1.11-1.25). (45) The hazard ratio for a 1 kg/m^2 unit increase in BMI was 1.02 in the current study. Folsom (60) reported a hazard ratio of 0.95 for BMI for both men and women, but their model included waist to hip ratio which may have led to a lower hazard ratio for BMI than would have been produced by a model with BMI alone.

There were few prior CHD models which produced comparable hazard ratios. Where they did there were some differences which could not be directly accounted for (age, HbA_{1C}): both of these were from the UKPDS model reported by Stevens. (45) However, overall, these comparisons suggest that the CHD prediction model created as part of the current study produced equivalent results to other prediction models based on UK data

Stroke: Two stroke risk prediction models specific to people with diabetes were identified from the literature (table 8.5). (46, 58) As with CHD no other model used the full set of covariates included in the current study model. The inclusion of CHD and CKD (outcomes which increase with age) as predictors, therefore, might tend to result in a lower hazard ratio for age itself in the current model compared with models which did not include them.

Yang (58) reported comparable results for age, prior CHD and HbA_{1C}, while Kothari (46) reported age, male gender and current smoking. The hazard ratio for male gender was positive (HR 1.15) but not statistically significant in the current study. Yang (58) did not report an estimate for the effect of gender indicating that it, like the current study, was not statistically significant. The equivalent result for Kothari (46) was 1.42 (95% CI 1.09-2.06). This 95% confidence interval was relatively wide and overlapped with the estimate from the current study, suggesting that the observed difference was non-significant.

The hazard ratio for age in the current study was 1.06. The equivalent result from the Yang model (58) was 1.07 (95% CI 1.06-1.08), and the result from the Kothari (46) model was 1.09 (95% CI 1.07-1.12). The estimate for the current study lay within the 95% confidence intervals reported by both Yang and Kothari.

The current study and the Kothari (46) stroke model produced similar estimates of the effect of current smoking. The hazard ratio for smoking in the current study was 1.42. This was within the 95% confidence intervals reported by the Kothari model (HR 1.55; 95% CI 1.08-2.01). The hazard ratio for prior CHD in the current study (HR 1.50) was also within the wide 95% confidence intervals reported by Yang (58) (HR 1.76; 95% CI 1.15-2.69).

A 1% increase in HbA_{1C} did appear to increase the risk of stroke to a small extent (HR 1.01) in the current study, but was not statistically significant. HbA_{1C} did not achieve statistical significance in the stepwise selection method used to develop the model reported by Kothari (46) either, but was significant and positive in the model reported by Yang (HR 1.09; 95% CI 1.02-1.08), although the lower boundary of the 95% confidence interval indicated that its importance may be small in clinical practice (a 2% increase in the risk of stroke for each additional unit increase in HbA_{1C}).

Overall, gender did not seem to influence the risk of stroke in the current study model: the results of the two other comparable models did not contradict this. (46, 58) A similar pattern was seen for age, smoking and HbA_{1C}: the effect of each of these was to increase risk and there was an overlap between the confidence intervals for the current and comparison studies suggesting that any observed differences were non-significant. This suggests that the current study model for stroke has face validity, at least for the set of predictors that could be compared.

CKD: Only three of the nine CKD prediction models identified in the literature search analysed their data as a time-to-event/ survival model (table 8.6). (67-75) The remaining six used a logistic model which does not produce directly comparable results. (67, 68, 70, 72, 74, 75) Of the three models which reported their results as hazard ratios (69, 71, 73), two reported at least one predictor which overlapped with the current study. (69, 73).

Male gender was associated with an increased risk of CKD in the current study and the prediction model published by Hanratty. (69) The hazard ratio was higher in the current study (HR 0.52) than in the Hanratty model (HR 0.63; 95% CI 0.59-0.66) and did not overlap with its 95% confidence interval. The effect of age was published by both remaining studies: the hazard ratio for each year increase in age in the current study was 1.06. This is the same as that reported by Hanratty (HR 1.06; 95% CI 1.05-1.07), and was slightly lower but overlapped the 95% confidence interval reported by Chien (HR 1.08; 95% CI 1.07-1.10). (73) The effect of prior CHD was similar in both the current study (HR 1.21) and Hanratty (HR 1.24; 95% CI 1.15-1.33). The effect of prior stroke in the current study (HR 1.14) was also similar to that reported by Hanratty (HR 1.08; 95% CI 0.95-1.22), and substantially lower than, but within the wide 95% confidence intervals reported by Chien (HR 3.46; 95% CI 1.27-9.38). The effect of the last common predictor, increasing BMI in the current study, like age,

was the same as that reported by Hanratty (HR 1.01 for both), and within the 95% confidence intervals reported by Chien (HR 1.12; 95% CI 1.01-1.12).

Despite the differences in the populations used in the current study and in the models used in these comparisons, similar results were seen for age, prior CHD and BMI. The sole study which produced comparable results for male gender was in the same direction but approximately 20% higher than the current study. This level of difference may be accounted for by other differences between the study populations and the predictors included in each model. Overall, this suggests that the current study model for CKD has face validity, at least for the set of predictors that could be compared.

All-cause mortality: As previously discussed, the results from the current study for mortality appeared to be the most robust. Six other all-cause mortality prediction models intended for use in the diabetic population were identified from the literature (table 8.7). (76-81) Two of these models had two predictors in common with the current study (79, 80), and a further two had just one predictor in common (HbA_{1C}). (76, 78) The Skriver model (77) reported results for HbA_{1C} as a set of categorical predictors rather than a single continuous one, and the Clarke model (81) published a logistic model for mortality, and so were not comparable with the current study.

Male gender was associated with a similar increased risk of death in the current study (HR 1.29) and the prediction models published by Kerr (HR 1.20; 95% CI 1.0-1.5) and Currie (HR 1.34; 95% CI 1.26-1.43), and lay between the 95% confidence intervals for each of these studies. (79, 80) The hazard ratio for each additional year of age in the current study was 1.09: this was very close to the ratio of 1.08 (95% CI 1.08-1.09) published by Currie.

The impact of smoking on the risk of death in the current study (HR 1.65) was lower than that reported by the Kerr model (HR 2.1; 95% CI 1.5-2.8) which, like the current study, included incident cases of type 2 diabetes. (79) The comparison group in the Kerr model was never-smokers, unlike the current study, where it was with ex- and current smokers. This and the overlap of the confidence intervals from the Kerr study with the current study estimate probably go some way to explaining differences in the hazard ratio estimate between the two studies.

The hazard ratio for a one unit increase in HbA_{1C} on the risk of death in the current study was 1.09, less than that reported for the Xu (HR 1.13; 95% CI 1.06-1.20) (76) and Andersson (HR 1.16; 95% CI 1.09-1.23) (78) models, but overlapped with their 95% confidence intervals.

Overall, the estimates reported for the current study model for all-cause mortality were equivalent to the other similar models for comparable predictors. (76-81) The one instance where they were dissimilar (smoking), may be explained by differences in the way smoking status was categorised. This suggests that this prediction model has face validity when compared with equivalent models, at least for this common set of predictors.

Summary: A total of 16 comparable models were identified from the literature: between two and six external models for each of the current study models. (44-46, 57-60, 69, 71, 73, 76-81) Only four of the 16 external model cohorts consisted of newly diagnosed cases of type 2 diabetes (45, 46, 79, 81); six used prevalent cases of type 2 diabetes (57-60, 77, 78, 80), two used prevalent type 1 and type 2 cases (44, 76), and three included people without diabetes (69, 71, 73).

The most commonly reported comparable predictors from these earlier studies were age, gender and smoking status. (45, 46, 57, 59, 60, 69, 73, 79, 80) Comorbidities at baseline

(CHD and stroke, but not CKD), HbA_{1C}, and BMI were reported for between three and six external models. (45, 58, 60, 69, 73, 76, 78, 81) No other studies reported results for the effect of systolic BP, cholesterol or drugs used in their treatment in a form which could be compared with the current study models, and none included the level and treatment of all these risk factors. Given the importance of blood pressure and cholesterol as risk predictors (table 7.11) and the prevalence of these treatments at diagnosis of diabetes (table 7.5), this suggests that the current models would also be more useful in clinical practice.

The results for the effect of age, gender and current smoking at baseline in the current study models were in the same direction and on a similar scale to the estimates published in earlier models. The effect of CHD and stroke as comorbidities was also similar to that reported by other models. The estimate for HbA_{1C} on the risk of CHD, stroke and all-cause mortality was lower than that reported by other studies: it lay within the reported confidence intervals in the comparison studies for all-cause mortality, but just outside those for CHD and stroke. Comparisons of the effect of BMI on risk were not as consistent as for other predictors: it was higher than the sole external model for CHD that reported a hazard ratio for BMI (Folsom) (60), but the same as or within confidence intervals of the two external models for CKD prediction. However this difference may be explained by the predictors included in each model: in addition to BMI, the Folsom model included waist to hip ratio. This may have led to a lower hazard ratio for BMI than would have been produced by a model with BMI alone and explain the difference between the results of the two models.

Overall, despite the many sources of variation between the models generated in the current study and external models, the effect of age, gender, smoking, prior CHD and stroke did not differ consistently. However, where they did differ, the size of the difference was relatively small and may plausibly be due to differences in the membership of their respective cohorts.

It was, unfortunately, not possible, however, to externally validate the current study hazard ratios for CKD as a comorbidity or the blood pressure and cholesterol-related predictors. These predictors (systolic BP, total cholesterol and their respective drug treatments) were entered into their respective models as categorical terms with treatment interaction. It would have been useful to validate these estimates as many are relatively large (a 7%-69% change in risk over 5 years) but were not statistically significant. It remains unclear, as a result, if they are likely to be useful predictors of future risk and it is unlikely that a future study could achieve substantially greater power to give a clinically useful, suitably precise, estimate for these predictors (as the current study used one of the UK's three large GP databases as the data source).

Although the four prediction models produced by the current study appear to have face validity following comparisons with prior prediction models, their external validity hasn't been established. It would be necessary to go beyond the scope of the current study to validate it using a second set of primary care data or by using it prospectively in clinical practice [section 8.7.2].

Table 8.4 Comparison of hazard ratios with other CHD prediction models specific to people with diabetes

Lead author (reference)	Study population	Hazard ratios											Predictors in model
		Male	Age	Smoker	Stroke	CKD	HbA_{1C}	BMI	SBP >= 140 mmHg or drug trt. BP	Total chol. >= 4 mmol/L	Total chol. >=4 mmol/L and on lipid lower.drug	On lipid lowering drug	
Current study	***Incident type 2 diabetes (age 35+)***	***1.36***	***1.02***	***1.26***	***1.09***	***1.13***	***1.07***	***1.02***	***1.41***	***1.38***	***0.98***	***1.79***	***(see column titles)***
Yang (57)	Prevalent type 2 diabetes (free of heart failure)	2.01	1.03	1.55	-	-	-	-	-	-	-	-	Age, diabetes duration, sex, smoking, eGFR, albumin:creatinine ratio, non-HDL cholesterol, total:HDL cholesterol ratio, HbA_{1C}, Systolic BP
Donnan (59)	Prevalent type 2 diabetes with complete risk factor data	0.73	0.97	0.76	-	-	-	-	-	-	-	-	Age at diagnosis, duration of diabetes, HbA_{1C}, smoking, sex, systolic BP, treated hypertension, total cholesterol, height
Folsom (60)	Prevalent type 2 diabetes, (aged 45-64 years) (men and women modelled separately)	-	-	1.05(m) 1.57(f)	-	-	-	0.95(m) 0.95(f)	-	-	-	-	Age, age-squared, race, smoking, total cholesterol, HDL cholesterol, systolic BP, use of antihypertensives, smoking status. BMI, waist:hip ratio, heart rate, physical activity, FEV, Keys score, tobacco pack-years, creatinine, albumin, factor VII, WBC, LVH, carotid IMT factor VIII, von Willebrand factor
Stevens (45)	Incident type 2 diabetes (aged 25-65, no recent history of CHD) (model data was from 4 years post diagnosis)	1.69	1.06	1.35	-	-	1.18	-	-	-	-	-	Age at diagnosis, sex, ethnicity, smoking, HbA_{1C}, systolic BP, total:HDL cholesterol ratio, duration of diabetes
Yudkin (44)	Prevalent type 1, type 2 diabetes (model coefficients/ hazard ratios not published)	-	-	-	-	-	-	-	-	-	-	-	Age, sex, smoking, microalbuminuria, total:HDL cholesterol ratio

Table 8.5 Comparison of hazard ratios with other stroke prediction models specific to people with diabetes

Lead author (reference)	Study population	Hazard ratios											Predictors in model
		Male	Age	Smoker	CHD	CKD	HbA_{1C}	BMI	SBP >= 140 mmHg or drug treated BP	Total chol. >=4 mmol/L	Total cholesterol >= 4 mmol/L and on lipid lowering drug	On lipid lowering drug	
Current study	Incident type 2 diabetes (age 35+)	(1.15)	1.06	1.42	1.50	(1.17)	(1.01)	(0.99)	1.80	(0.90)	(1.22)	(0.63)	(see column titles)
Yang (58)	Prevalent type 2 diabetes	n/s	1.07	n/s	1.76	-	1.09	-	-	-	-	-	Age, HbA_{1C}, albumin:creatinine ratio, CHD, sex, smoking, SBP, total:HDL cholesterol ratio
Kothari (46)	Incident type 2 diabetes (aged 25-65, no recent or multiple CHD events, followed-up from 4 years after diabetes diagnosis)	1.42	1.09	1.55	-	-	-	-	-	-	-	-	Age at diagnosis, duration of diabetes, sex, smoking, systolic BP, total:HDL cholesterol ratio, atrial fibrilation

Table 8.6 Comparison of hazard ratios with other CKD prediction models for wider population

Lead author (reference)	Study population	Hazard ratios											Predictors in model
		Male	Age	Smoker	CHD	Stroke	HbA_{1C}	BMI	SBP >= 140 mmHg or drug trt. BP	Total chol. >= 4 mmol/L	Total chol. >= 4 mmol/L and on LLD	On lipid low. drug (LLD)	
Current study	***Incident type 2 diabetes (age 35+)***	***0.52***	***1.06***	***(1.10)***	***1.21***	***1.14***	***1.03***	***1.01***	***1.45***	***(1.46)***	***(0.64)***	***(1.69)***	***(see column titles)***
Hanratty (69)	Hypertensive adults	0.63	1.06	-	1.24	1.08	-	1.01	-	-	-	-	Age, gender, race/ethnicity, baseline eGFR, SBP, HDL cholesterol, BMI, diabetes, CHD, CVD, heart failure, PVD
Chien (73)	Wider population	-	1.08	-	-	3.46	-	1.06	-	-	-	-	Age, BMI, diastolic BP, type 2 diabetes, stroke
Hippisley-Cox (71)	Wider population (ages 35-74) (fractional polynom. or categorised values reported)	-	-	-	-	-	-	-	-	-	-	-	Age, ethnicity, deprivation, smoking, BMI, SBP, diabetes type, rheumatoid arthritis, CVD, treated hypertension, congestive cardiac failure, PVD, NSAID use, FH of kidney disease, systemic lupus erythematosis, kidney stones

Studies where a logistic model was used to predict future risk (no direct comparisons with hazard ratios from current study possible)

Lead author (ref)	Study population	Predictors in model
O'Seaghdha (67)	Wider population (ages 30-62)	Age, diabetes, hypertension, baseline estimated glomerular filtration rate, albuminuria
Alssema (68)	No type 2 diabetes, no CVD (ages 25-85)	Age, smoking, use of antihypertensives, use of lipid-lowering drugs, BMI, waist circ., FH <65 years of MI/stroke, FH diabetes, hist. of gest. diabetes
Halbesma (70)	Wider population (ages 28-75)	Baseline eGFR, age, urinary albumin excretion, systolic BP, C-reactive protein, and known hypertension
Hanratty (72)	Hypertensive adults (age 21+)	Age, sex, race/ethnicity, marital status, language, diabetes, vascular disease, heart failure, dyslipidaemia, major psychiatric diag., substance abuse, baseline eGFR
Kshirsagar (74)	Wider population (ages 45+)	Age, sex, race/ethnicity, anaemia, CVD, diabetes, heart failure, PVD, HDL cholesterol
Fox (75)	Wider population (ages 30-62)	Age, sex, baseline eGFR, BMI, smoking, diabetes, systolic BP, hypertension, hypert. treatment, total chol., HDL, impaired fasting glucose

Table 8.7 Comparison of hazard ratios with other all-cause mortality prediction models specific to people with diabetes

Lead author (reference)	Study population	Hazard ratios												Predictors in model
		Male	Age	Smoker	CHD	Stroke	CKD	HbA_{1C}	BMI	SBP >= 140 mmHg or treat. BP	Total chol. >= 4 mmol/L	Tot. chol. >= 4 & on LLD	On lipid low. drug	
Current study	***Incident type 2 diabetes (age 35+)***	***1.29***	***1.09***	***1.65***	***1.60***	***1.47***	***1.33***	***1.09***	***0.98***	***(1.07)***	***(0.61)***	***(1.69)***	***0.41***	***(see column titles)***
Kerr (79)	Incident type 2 diabetes	1.20	-	2.10	-	-	-	-	-	-	-	-	-	Age group, sex, year of diagnosis, HbA_{1C} category at 3 months, systolic BP, smoking
Currie (80)	Prevalent type 2 diabetes, on combination oral antidiabetic treatment or insulin (aged 50+)	1.34	1.08	-	-	-	-	-	-	-	-	-	-	Age, sex, smoking status, cohort (combination therapy or insulin initiated), HbA_{1C}, mean total cholesterol, LVD, Charlson Index
Xu,(76)	Prevalent type 1 and type 2 DM (age 65+)	-	-	-	-	-	-	1.13	-	-	-	-	-	Age, sex, education, smoking, alcohol use, exercise, CVD, BMI, total cholesterol, HbA_{1C}
Andersson (78)	Prevalent type 2 diabetes, high BMI and high CVD risk	-	-	-	-	-	-	1.16	-	-	-	-	-	Age, sex, randomised treatment assignment (sibutramine), diabetes duration, history of: arterial hypertension/congestive heart failure/CVD/revascularisation, ethnicity, tobacco use, SBP, DBP, heart rate, HbA_{1C}, BMI, HDL chol., LDL chol., urine albumin/creatinine ratio and use of insulin, metformin, thiazolidinediones and sulfonylureas
Skriver (77)	Prevalent type 2 diabetes (HR for HbA_{1C} categories reported only)	-	-	-	-	-	-	-	-	-	-	-	-	Age, sex, diabetes duration, mean annual HbA_{1C} at baseline, CVD, arteriosclerosis, acute complication of diabetes, retinopathy, nephropathy, MI, stroke, neuropathy

Study where a logistic model was used to predict future risk (no direct comparisons with hazard ratios from current study possible)

Lead author (ref)	Study population	Predictors in model
Clarke (81)	Incident type 2 diabetes (aged 25-65)	Age, sex, smoking status (ever vs never), HbA_{1C}, total:HDL cholesterol ratio, MI, renal failure, amputation

8.5 Other analysis methods

Understanding the relevance of the present study requires comparison to the methods used elsewhere as well as the comparative results. Sixteen of the 23 studies identified from the literature used time-to-event statistical models like the current study. (44-46, 57-60, 69, 71, 73, 76-81) The remaining seven used logistic models. (67, 68, 70, 72, 74, 75, 81) The use of time-to-event models for the current study was appropriate for the source data, and better than the use of logistic models, given the censoring of outcomes when patients move practice. (103) It allowed patients who left their practice before the end of the five-year follow-up to be included in the analyses, and the observed survival time to be used in place of a simpler, less informative, binary outcome. As a result the time-to-event models made better use of the available data, and were powered to estimate the effect of each predictor with greater precision than their logistic counterparts in the literature.

Only four of the 13 previous prediction models, which were specific to diabetes, modelled risk from the point of diabetes diagnosis (incident cases). (45, 46, 79, 81) The remaining nine included cases who had been diagnosed at some time in the past (prevalent cases): the level of their risk factors were estimated at the time of their entry into follow-up, rather than at the diagnosis of diabetes. (44, 57-60, 76-78, 80) Where diabetes duration was included as a predictor in four of these nine models, it was entered as a single covariate, with no interactions between it and other covariates. (57, 59, 77, 78) Data from the current study show how high levels of HbA_{1C} (figure 7.3) and cholesterol (figure A7.2) at diagnosis of diabetes declined substantially in the first few years as treatment was initiated. Models that derived their HbA_{1C} and cholesterol data from prevalent cases of diabetes did not make any allowance for this feature of the data (at a minimum by including an interaction term for HbA_{1C}

/cholesterol and time since diagnosis). (44, 57-60, 76-78, 80) If used with newly diagnosed case of diabetes, these models are likely to overestimate the effect of HbA_{1C} and total cholesterol on outcomes [section 8.3.1, paragraph 4 and section 8.9.1 paragraph 4]. This is consistent with the somewhat higher hazard ratios for HbA_{1C} reported by these studies and reinforces the clinical utility of the current study (tables 8.5 and 8.7). (58, 73, 76)

A similar effect may also be present in the UKPDS-derived CHD prediction model which used incident cases of diabetes and the average of HbA_{1C} at one and two years following diabetes diagnosis (table 8.4) (45): they also reported higher hazard ratios for HbA_{1C} than were estimated in the current study CHD model. This implies that the current study models, with their inclusion of clinical values at diagnosis of diabetes, would be more accurate estimates of future risk when used with patients with newly diagnosed diabetes than previous models. It could be a useful tool for use with patients with newly diagnosed diabetes to identify and target preventative treatment at those with highest risk, and to advise these patients of their likely prognosis.

The use of risk factor levels at baseline to predict future risk over relatively short periods (five years) may be appropriate for predictors which are subject to treatment (HbA_{1C}), or more intensive treatment (cholesterol and blood pressure). However, if the follow-up period was to be extended to more than 10 years in future models, account would need to be taken of regression dilution, where exposure levels at baseline do not reflect the relationship between exposure levels in later periods and subsequent risk. (218) For shorter follow-up periods baseline risk levels may accurately reflect the accumulated exposure to the risk factor, than measurements from later periods. Future models that follow cases from the point of diagnosis for periods of 10 years or more may, therefore, benefit from the use of multiple measurements for each of these risk factors: a single measurement at diabetes diagnosis which reflects past

(untreated or less aggressively treated) exposure, and one or more additional measurements separated by several years to identify and account for any regression dilution effect. (218)

8.6 Other approaches to missing data

8.6.1 The strengths and weakness of the approach used in the current study

The extent of missing data in this study meant that care had to be taken in selecting appropriate methods for dealing with it. Relatively few (18%, 3643/20041) of the cases in the study cohort had all their clinical measurements recorded (that is, coded in their electronic patient record) within 90 days of diabetes diagnosis (table A7.1). After modelling their level using later values, baseline clinical measurement data was still missing for between 1% and 7% of cases [section 7.4]. The current study used two methods in combination to deal with missing baseline clinical measurements (HbA_{1C}, systolic BP, total cholesterol, BMI, eGFR/creatinine) and Townsend deprivation quintile. This involved estimating the baseline level of clinical measurements for cases using data from later time periods using a set of multilevel models [section 6.6.7] and using multiple imputation to estimate the level of the missing data where it could not be estimated using these multilevel models [section 6.7.5]. It had the following strengths and weaknesses.

Strengths: (a) The proportion of cases with missing clinical values was minimised by thorough data cleaning and inclusion of data recorded as free text. (b) Patients' own data were used to estimate their baseline clinical value levels, rather than treating it as missing and imputing from the values recorded from complete cases. (c) The multiple imputation process made an allowance for the imprecision of its estimates which was reflected in the confidence intervals for the hazard ratios for clinical values in the final prediction models.

Weaknesses: (a) The free text search (looking for clinical values added as text) was time-consuming and did not identify a significant number of new values [section 7.4]. (b) The model to estimate baseline systolic BP did not produce very accurate estimates of baseline values (figure 7.4): systolic BP had to be entered into the survival models as a binary rather than a continuous predictor as a consequence of this [section 7.6.1]. This reduced the ability of the outcome models to identify the relationship between this predictor and each of the outcomes of interest.

Only 18% of patients with newly diagnosed type 2 diabetes between 1998 and 2003 had all their clinical measurements recorded within 90 days of diagnosis (table A7.1) It is possible that the recording of these values in electronic patient records close to the time of diagnosis has improved following the introduction of electronic linkages to laboratories and QOF, although it appears that recording did not improve for newly diagnosed cases as much as for prevalent diabetes cases in the period up to 2007. (12, 103) This suggests that missing baseline data would still be an issue if the models presented here were used in current clinical practice. Practices would have to either impute missing values or measure them directly when using these models to estimate risk for an individual. This would be possible in an individual patient setting, where BP could be measured during the consultation, but it might require a new blood test to estimate total cholesterol or eGFR, increasing the burden on practice staff and delaying the risk assessment. Alternatively, these data may already be to hand in the form of a scanned hospital letter, in which case they would only need to be entered by hand into the appropriate section of the electronic patient record.

An alternative to using actual measured values from an individual would be to impute missing values using estimates from the wider population or from the cohort used to derive the prediction models. (219, 220) This would also be required if estimating risk for a large group

of individuals simultaneously. The current models do not yet provide a means of doing this, but one could be incorporated in the software used to estimate risk at a later stage. Whatever the final source of data for these imputations, it would be valuable if the uncertainty surrounding the precision of a risk estimate derived for an individual patient using imputed data could be reported by the clinical software. This could be reported as an upper and lower estimate of risk which would be wider for a patient where one or more values were imputed, and narrower where none were missing.

8.6.2 The approach used in earlier prediction models

Ten of the 23 prediction models reported using at least one method for dealing with missing predictor data. (45, 46, 59, 60, 71, 72, 78-81) One of them reported using three separate methods for handling missing data. (71) The most common method used was complete-case analysis. (45, 46, 59, 60, 78, 79, 81) Aside from reducing the power of a prediction model, this approach may cause bias and is not recommended unless the proportion of missing data is low [section 6.7.5]. (99, 221)

Single imputation: Single imputation (using a population mean value or a default category) was used in two studies.(71, 72). The multilevel models used in this study estimated baseline values more accurately than the mean of the observed values for each case [section 7.4], suggesting that this method was preferable to mean imputation. It would not have been appropriate to use the alternative single imputation method (a default category) (71) in this study as it would have assumed that all cases from the current study with missing Townsend deprivation quintile could safely be assigned to a single default quintile as the population as a

whole was relatively evenly distribution across deprivation levels. This approach would have assigned cases with missing Townsend to the quintile with the greatest number of cases, and led to an systematic underestimation of the effect of deprivation on outcomes, and overestimated the precision of these estimates.

Last value carried forward / last value carried back: Last value carried forward was used in one (80) and the closest in time to baseline of last value carried forward and last value carried back was used in a second prediction model identified from the literature. (71) It may be unsafe to use these simple approaches as they treat risk factors recorded at different times, perhaps years apart, as if they were all recorded on the same date and not correlated with one another. It also assumes that there was no risk of recording bias. For example, some risk factors may be first recorded as part of the investigation of an outcome of interest, for example a family history of gastrointestinal cancer or higher levels of alcohol use may be first recorded when cancer is suspected. (101) The inclusion of these records, close to the cancer diagnosis, would tend to overestimate the relationship between family history/alcohol use and the risk of that cancer.

The multilevel models used in this study were more likely to produce accurate estimates than these two other simple methods as: (i) they took into account the relatively large changes in the levels of some risk factors following diabetes diagnosis (figures 7.3, 7.4 and A7.2; and (ii) the level of some risk factors would not have been routinely measured prior to diabetes diagnosis (HbA_{1C}, creatinine) and were not part of routine screening in the healthy population (total cholesterol). The implication of the approach to missing baseline values adopted in this study is that the values entered into the analyses used to produce the prediction models were more likely to represent their true value at baseline than other approaches. These values are

more likely to be close to the actual baseline levels experienced by patients at diagnosis, and therefore produce more accurate estimates of future risk in clinical practice.

Multiple imputation: Multiple imputation, as used in the current study and one of the other models (71) is currently regarded as an appropriate approach to missing covariate data in predictive models. (50, 99, 143) It does rely on the assumption that data are missing at random: this may hold for risk factors where practice patients would be expected to have this recorded routinely (e.g. total cholesterol in patients with type 2 diabetes), but may not hold for the healthy population (i.e. where cholesterol screening is not routine). However, the use of multiple imputation in the latter situation did not appear to have systematically under- or overestimated total cholesterol levels in those with missing data in one CVD prediction model based on the healthy population. (64)

8.6.3 Approaches which may be available for future studies

More efficient approaches to missing baseline data would remove the need to fill them in using two separate processes, but none have yet emerged. Standard multiple imputation, as used in this study, uses non-missing observations from others to impute an individual's missing data, and does not make use of that individual's non-missing observations in other time periods. The two-fold fully conditional specification algorithm appears to offer a partial answer to the need to estimate baseline levels of risk factors, but requires that individuals have a recorded value before and after baseline. (222) These will not always be available, particularly for risk factors which are recorded more frequently following diagnosis of diabetes (e.g. HbA_{1C}, total cholesterol), and so only provides a partial solution. Ongoing

work, led by University College London, to develop imputation models for missing data in primary care databases may provide a more comprehensive approach like that outlined above. (223) This could be used in future prediction models like those in this study.

8.7 Clinical implications

8.7.1 What risk factors make most difference and what opportunities are there for risk to be reduced?

The results presented in table 7.11 show the relationship between each of the risk factors and the risk of CHD, stroke, CKD and all-cause mortality. Modifiable risk factors are considered here first.

Smoking: Smoking was common and had a significant influence in all of the models: one-quarter of patients were smokers at baseline (table 7.5), and half of these still smoked at five years following diagnosis (table A7.2). The risk of all four outcomes was higher in smokers (65% higher in for all-cause mortality) (table 7.11), suggesting that intensive smoking cessation interventions would be appropriate in aiming to reduce the risk of all four outcomes in individual patients.

BMI: BMI at diagnosis was high for a large proportion of cases (median BMI=29 kg/m^2) (table 7.5) and did not change substantially over the five years following diagnosis (figure A7.1). Higher baseline BMI increased the risk of CHD and CKD significantly: a 3.18kg (7lbs/ half-stone) increase in weight was associated with a relatively modest 3% and 1.5% increase in the 5-year risk of CHD and CKD, respectively (table 7.11). The greatest reductions in risk may, therefore be obtained by the 25% of patients whose baseline BMI was in excess of 33 kg/m^2. Given that weight loss is associated with a lowering of CVD risk and better glucose control (26, 29), this suggests that there are significant further opportunities to reduce risk by interventions aimed at weight loss, for example early referral to patient education programmes. (224)

HbA_{1C}: Higher HbA_{1C} at diagnosis of diabetes was a statistically significant predictor for CHD and CKD, and highest for all-cause mortality (9% risk increase for each 1% HbA_{1C} increase) (table 7.11). Following diagnosis, mean HbA_{1C} decreased significantly in the subsequent six months among the groups with the highest levels at diagnosis (HbA_{1C} 9.5%+ and HbA_{1C} 7.5%-9.4%) (figure 7.3). A similar pattern was seen in the UKPDS (figure 7.5), though their data were reported from one year following diagnosis rather than diabetes diagnosis itself. (81) These early reductions in the current study were maintained in the following five years, but there was a small and observable annual increase in all HbA_{1C} groups from one year following diagnosis (figure 7.3). By five years, only the group with the lowest initial HbA_{1C} (baseline HbA_{1C} of under 6%) remained under the current NICE target of 6.5%. (43) This indicates that initial improvements in HbA_{1C} control were not followed up by further successful attempts at control after the first year and that more aggressive and sustained blood glucose control may be indicated for those with higher baseline HbA_{1C} (i.e. those with HbA_{1C}>6%: the majority of cases). Further reductions in HbA_{1C} in the period after the first year following diagnosis should lead to reductions in the risk of major outcomes. (28)

Systolic BP: Hypertension at diagnosis (high systolic BP or treated BP) was associated with an increased risk of CKD and CHD and was highest for stroke (an 80% increase in risk) (table 7.11). This was similar to that observed in the literature, where BP was a more important risk factor for stroke than CHD. (40) More than half of patients had a high SBP (>=140 mmHg) at baseline (table 7.5). Average systolic BP slowly declined in this group over the follow up period but was still high at the end of the fifth year (figure 7.4). A similar progression can be seen in data reported by the UKPDS (figure 7.5). This indicates that there is opportunity for improved management of this risk factor from diabetes diagnosis to at least five years, which should have the effect of reducing the risk of the above diabetic complications. (27, 29, 32,

34) Such reductions would also impact on microvascular disease and hence are doubly important. (225)

Total cholesterol: High total cholesterol at diagnosis (>=4 mmol/L) was associated with an increased risk of CHD and CKD, but was not statistically significant (table 7.11). More than three-quarters of patients had high total cholesterol at baseline (25^{th} percentile 5.1 mmol/L) (table 7.5). Average total cholesterol levels did decline over the five-year follow-up for patients with a baseline cholesterol of 5 mmol/L or higher, but did not fall below the cutoff of 4.5 mmol/L by the end of this period (figure A7.2). Table A7.3 also shows that the prevalence of lipid-lowering drug use increased from 19% of patients at baseline to 42% at one year following diagnosis, but does not provide any data past this point. However, the high total cholesterol observed at five years in those with the highest baseline levels (figure A7.2) indicates that there was scope for further reductions in cholesterol levels and, therefore, risk of diabetic complications from improved drug and lifestyle changes. (27, 29, 32, 34)

Prior comorbidity: Prior comorbidity (CHD, stroke or CKD) at diagnosis of diabetes was relatively common: more than one in five patients had at least one comorbid condition (table 7.5). These also had a large effect on the risk of death in the first five years (table 7.11). Patients with two of these comorbidities had approximately twice the risk of death as patients free from them. Although not modifiable, the increased risk of major outcomes following diabetes diagnosis should be recognised in this group: these patients could be flagged for intensive treatment to reduce risk. (38, 39) In addition to this, a substantial number of additional patients (5%) were found to have CKD in the three months following diabetes diagnosis (table 7.1), presumably when their creatinine level was first measured as part of routine diabetes care. (39) This suggests that there may be a benefit in checking creatinine earlier, possibly when patients are being assessed for diabetes or in patients with pre-diabetes,

as it would allow treatment to preserve kidney function to be initiated at an earlier stage in the disease.

Given the high risk of CVD, CKD and death in patients with prior comorbidity, an argument can also be made in favour of routinely screening these patients for diabetes [section 8.8 policy implications: clinical issues].

8.7.2 What is the clinical utility of the risk prediction models developed in this study?

The clinical utility of the models presented in this thesis is assessed here in terms of how they could be used in clinical practice to improve health outcomes. The final part of this section describes a possible use of the models in a likely clinical setting, once they were validated and improved in the manner suggested below.

Broadly, a clinically useful model should: predict the risk of an important health outcome; provide thresholds for action; trigger the use of available and safe interventions to reduce risk; and be cost-effective. (226) In addition to helping clinicians make treatment decisions about individual patients, risk models may help patients understand their risk of disease and motivate them to initiate behavioural change or improve adherence to prescribed treatments. (227) Risk prediction models may also be used at a population level, to allocate resources to those at highest risk [section 1.3]. (38, 105, 228, 229)

The models presented in this thesis predicted the risk of three important outcomes, namely CVD, CKD and death. Clinically- and cost-effective interventions to reduce these risks are available in UK general practice - the likely setting where these models would be used [section 1.2]. The risk of these outcomes is higher in people with type 2 diabetes [section 1.4]

but could be reduced by lifestyle changes and medical management [section 1.2] which are relatively cheap (as generically prescribed drugs) and available (e.g. the DESMOND patient education programme) to patients in the UK. (230) It is also possible that better understanding of prognosis on the patients' part might lead to more effective implementation of such interventions. (200-202, 204, 231)

However, the models presented here do not currently have thresholds for action (e.g. prescribe statin/ BP lowering drugs to all patients with greater than 20% 5-year risk of CVD or death), and their cost-effectiveness, if they were used in UK general practice, is not known. Further work, beyond the scope of this thesis, would be required to validate their clinical utility and impact in these particular respects. (232-235)

Simply reporting the limitations of previous attempts to predict risk in people with type 2 diabetes and the potential superiority of the new models, described earlier [sections 1.4, 2.3, 2.4, 3.3], is insufficient to recommend their adoption at this stage. The models do provide estimates of absolute risk, and the calibration plots presented in section 7.8 and the proportion of variation (R^2) statistics [section 7.7] suggest that the models may perform at least as well as other risk models currently used in UK clinical practice. (234) Additional research would also be required to quantify (e.g. measures of discrimination, calibration, and (re)classification) and externally validate their performance prior to adoption for use in clinical practice. (232-234, 236, 237)

An early use of the models in general practice might be to rank newly diagnosed type 2 diabetes patients by risk of CVD/CKD and death so that those at the highest risk could be allocated early assessment or additional interventions to manage their risk factors. This would be an alternative approach to the current incentives to general practices which focus on

individual risk factors. (84) The risk models would be relatively straightforward to implement in current clinical systems by third parties or the system suppliers, using the coded clinical data that they contain to populate the risk equations: this has been done elsewhere. (104, 238) The cost of making these models available in all UK practices using each clinical system would, therefore, not be significantly higher than installing it in a single practice. In such an application, the models would only to be required to discriminate between low and high risk patients. (239) The risk of misclassification for any individual patient (a high risk individual being misclassified as low risk or vice versa) would also be minimised in this application, as individual risks factors would eventually be addressed in routine care, probably at the time of their diabetes annual review. CVD and CKD risk are known to increase with age: older patients with newly diagnosed type 2 diabetes may already be on appropriate treatments to reduce these risks, or may be initiated on treatment without having their individual risk calculated. These models may, therefore, be most useful in younger patients to stratify and treat them according to their overall risk, rather than the level of their individual risk factors.

8.8 Policy implications

The previous section indicated that there were additional clinical opportunities available to reduce the level of modifiable risk factors in the years following diabetes diagnosis [section 8.7.1] and discussed their clinical utility [section 8.7.2]. The identification of future CVD, CKD and mortality risk among patients with newly diagnosed diabetes, and more aggressive management of risk factors in the years immediately following diagnosis should result in improved health outcomes [sections 1.3, 1.4]. The prediction models created in this study may prove particularly useful for this task as they were developed for this specific patient group [section 2.5], use risk factors which should be routinely available at diabetes diagnosis [section 4.3], and predict the risk of important clinical outcomes [section 1.2]. This section describes the implications of these and the other results presented in this thesis for policy makers. These are separated into two subsections: clinical issues and data issues.

Clinical issues: In addition to the opportunities available to reduce the level of modifiable risk factors in clinical practice in the years following diabetes diagnosis [section 8.7.1], there was also a high level of co-morbidity at diabetes diagnosis: unrecognised CKD along with previous stroke and CHD were frequently present at baseline (table 8.2). Assuming that these trends continued past the end of the study period (circa 2004) and are still common among patients diagnosed in more recent periods, approximately 10 years since the introduction of QOF, then there may be scope for additional national guidelines or incentives specific to patients newly diagnosed with Type 2 diabetes. This may be true even if the management of prevalent Type 2 diabetes appears to have improved since the introduction of QOF as the

management of this subgroup of newly diagnosed patients may not have improved to the same extent. (12, 13, 240)

The high levels (5%) of previously unrecognised CKD found in the first three months following a diagnosis of type 2 diabetes (table 7.1), and the fact that progression to CKD stages 3-5 occurred in one-third of patients overall, may suggest that targeted routine screening for CKD should be carried out at an earlier point, for example in patients with impaired glucose tolerance/impaired fasting glucose. Earlier identification of patients at risk of CKD and CKD itself could lead to earlier treatment initiation, delay further decline in kidney function, and reduce the risk of end-stage renal failure. (100)

A comparison of the trends in HbA_{1C} following diagnosis between the UKPDS (1977-1991) and the current study (1998-2003) for the first five years following diabetes diagnosis appears to show that blood glucose management had improved greatly, even before the introduction of financial incentives for diabetes management as part of QOF in 2004 (figures 7.3, 7.5). (13)

There was little change in BMI in the five years following diagnosis, irrespective of baseline BMI (figure A7.1) [section 8.7]. This suggests that there are significant further opportunities to reduce risk by weight reduction, for example utilizing early referral to patient education programmes such as DESMOND which can offer support and advice on weight loss and other diabetes-relates issues to newly diagnosed patients. (230)

Pre-existing CHD and stroke were relatively common at diabetes diagnosis and were strong predictors of future stroke, CKD and mortality (table 7.11). Given the existence of effective treatments, this suggests that these patients should be targeted for intensive treatment, irrespective of the level of their other clinical risk factors. (39) An argument can also be made in favour of routinely screening these patients for diabetes. This would allow their diabetes to

be identified at an earlier stage and damage associated with prolonged exposure to high blood glucose levels to be prevented through earlier intervention.

Data issues: The current study shows the importance of reporting a full set of covariates. Of the six previous all-cause mortality prediction models identified in this thesis (table 2.3), three carried out multivariate prediction models but only two reported results for the effect of HbA_{1C} on mortality (table8.6). A further single model to predict CHD risk reported its results as coefficients, rather than more easy to interpret hazard ratios (table 8.4). (59) When transformed into hazard ratios, male sex, increasing age and smoking could be seen to reduce the risk of future CHD, rather than to increase the risk as was observed by all the other models reviewed. (current study) (45, 57, 60) The reporting of full model results in a transparent format can allow flawed data or analyses to be identified and corrected. (64, 147) Reporting recommendations for predictive models should include reporting of model results for the full set of covariates and reporting those results in a transparent form (e.g. HRs not coefficients) in order to avoid selective non-reporting of results which do not fit previously published results.

Complex prediction models with a wide range of clinical predictors, which may include some with missing data, do not necessarily result in better or more useful models than simpler alternatives. The inclusion of predictors with missing data in this study led to complex data preparation and analysis (baseline prediction models and multiple imputation), that may not have been radically better than simpler demographic models. The availability of existing large UK primary care databases models makes the development of predictive models easier as they contain sufficient patients and outcomes to power a wide range of models. However,

they may appear to produce models which predict risk in UK populations better than models that are based on old or foreign data. This improved performance may only be because they have up to date data on the underlying risk in the population than the alternative sources. Reporting recommendations for predictive models should therefore report a summary statistic like R^2 or similar statistics for a basic and full model to allow the absolute value of the additional predictors derived from clinical data to be assessed.

A recent BMJ paper on the validity of acute MI diagnoses recorded in primary care suggests that these diagnoses have a positive predictive value of only 92% when compared with a disease registry data and that perhaps 25% of diagnoses were missing from primary care. (112) The authors conclude that linked primary care, death certification, hospital and disease registry data are required to avoid biased ascertainment of acute MI outcomes. These kinds of external data were not available at the time the models in this thesis were developed, but, as mentioned above, linked data is being introduced for use with THIN primary care records. (107) Future studies could make use of these data to validate existing GP data or identify additional outcomes [section 8.2.2].

8.9 Overall conclusions

Routinely collected primary care data can be used to predict future risk of coronary heart disease, stroke, chronic kidney disease, and all-cause mortality in people with newly-diagnosed type 2 diabetes mellitus. This last section of the discussion summarises the information that supports this main conclusion. It highlights the reasons why the current models may be more valid and clinically useful than previous models in a UK general practice setting. This includes their applicability to newly diagnosed patients with type 2 diabetes, the inclusion of predictors routinely available in general practice, the ease with which they can be updated in the future, and their management of missing baseline data. The section continues with a description of what might be usefully be communicated to patients, based on the study results, and ends with a summary of the study's scope, key findings, clinical utility, and recommends appropriate next steps in model development and validation.

8.9.1 How the prediction models developed in this study differ from past models

They are more applicable to newly diagnosed cases of type 2 diabetes: More than 20 models, identified from the literature, can be used to predict the risk of future CVD, CKD and death in people with type 2 diabetes (tables 2.2 to 2.3). Prediction models for CVD and all-cause mortality specific to people with diabetes which were developed using prevalent cases of diabetes are likely to overestimate risk when used with newly diagnosed cases of type 2 diabetes as the levels of HbA_{1C}, systolic blood pressure and total cholesterol decline in the years immediately following diagnosis [section 7.4]. Other models, developed using incident

(newly diagnosed) cases of diabetes were either derived from the UKPDS cohort and can only be used from four years following diabetes diagnosis (45, 46), or included year of diagnosis as a predictor and, therefore, can only be used with cases diagnosed between 1999 and 2007 (79). Some (for CKD risk) were intended for use with the general population (table 2.2): where these included diabetes as a predictor, they assumed that the effect of diabetes did not depend on the level of other risk factors, that duration of diabetes did not influence risk, and that diabetes control (HbA_{1C}) did not influence risk. (67, 69, 71-75)

This implies that the current study models, with their inclusion of clinical values at diagnosis of diabetes could provide more accurate estimates of future risk when used with patients with newly diagnosed diabetes than previous models. Of the four models presented, the most useful might be the all-cause mortality prediction model which appeared to explain a large proportion of the variability in clinical outcomes (R^2=0.58) This model, and the stroke and CKD prediction models, could be useful tools for use with patients newly diagnosed with diabetes: to identify and target preventative treatment at those with highest risk; and to advise patients of their likely prognosis.

They include a set of predictors available in general practice and can be easily updated: This study developed four separate prediction models which can be used with recently diagnosed cases of type 2 diabetes. These models predict the risk of CHD, stroke, CVD and all-cause mortality up to five years following diabetes diagnosis. They include a range of demographic and clinical predictors including some clinical measurements which are routinely assessed in patient with type 2 diabetes (HbA_{1C}, eGFR). These models can be easily updated to include cases diagnosed since 2004 (the end of the study period) using more recent primary care data, as has been done for the QRISK model. (109) This would allow the models to be extended to predict outcomes up to 10 years following diagnosis, and would provide

more precise estimates of risk in the first five years. Updates to these models could also include additional predictors not included in this study.

They handled missing data better than existing prediction models: The prediction models identified from the literature used a variety of approaches to handling missing baseline clinical measurements [section 8.6.2]. This included methods which may have introduced bias: complete-case analysis, single imputation using population mean values, and last value carried forward or back. (50, 145, 149, 221) One model which predicted risk in the general population did use a recommended approach, multiple imputation, but it was used in conjunction with last value carried forward/back. If this approach was used to develop a risk prediction model using data from patients with newly diagnosed type 2 diabetes it would systematically underestimate baseline risk levels: most of these values would be carried back from periods months after the diagnosis of diabetes, after lifestyle changes or new drug treatments had been initiated [section 6.6.7, table 7.4, table A7.1].

The approach adopted in this study was to estimate these baseline clinical measurements using multivariate models and then to use multiple imputation to fill in baseline values which could not be estimated [section 7.4]. This avoided introducing bias from the use of complete case, single imputation and last value carried forward/back approaches, and provided more accurate baseline estimates than single imputation and last value carried forward/back (table A7.1). The values that were multiply imputed at the next step in the process were probably also more accurate as a result. This process can be used in other studies where accurate baseline values need to be estimated, and can be used with longitudinal records where there are no data available prior to baseline, unlike more recent approaches such as the two-fold fully conditional specification algorithm. (241)

8.9.2 What to tell patients

Some of the risk of these major complications of diabetes cannot be altered (e.g. prior comorbidities), but their impact can be lessened by early interventions which reduce risk, such as drug treatments and lifestyle changes. (23-35) For all patients at high risk, there may be initial reductions in important risk factors such as HbA_{1C}, cholesterol and blood pressure once treatment is started, but they may well need to be followed up by further aggressive treatment to achieve and sustain treatment targets and reduce the risk of complications. Lastly, smoking is known to be associated with an increased risk of CVD and death among people with and without diabetes. (34, 35, 81) Because people with diabetes are at a higher risk of these and other significant outcomes, it is important to reduce as many risk factors as possible at the same time: this includes smoking. However, over half of people who smoked at diagnosis are still smoking five years later, so they may need increased support to quit and remain non-smokers.

8.9.3 Study summary

Routinely collected primary care data can be used to predict future risk of coronary heart disease, stroke, chronic kidney disease, and all-cause mortality in people with newly-diagnosed type 2 diabetes mellitus. This thesis developed four models which could be used to predict the risk of these outcomes in the five years following a diagnosis of type 2 diabetes. They may predict risk more accurately in the years following diabetes diagnosis than existing models: these were either developed using risk factors recorded some years after diagnosis,

included risk factors not routinely recorded in general practice or excluded important risk factors which are routinely recorded, or were developed for use in the wider population.

This study used data from a large UK general practice database and included risk factors which are known to predict these outcomes to populate the models: demographic variables, clinical predictors routinely recorded following diabetes diagnosis, and blood pressure and cholesterol-lowering treatment. Some of these models could, therefore, be used in a general practice setting to identify and target preventative treatment, and as educational tools to advise people newly diagnosed with type 2 diabetes of their likely prognosis.

Across models, the key modifiable predictors identified were: smoking; weight; blood pressure; and glycaemic control. The most clinically useful model might be the mortality model as it accounted for a large proportion of the variability in outcomes (R^2=0.58). This model found that age, sex and past medical history were associated with the risk of death, as were smoking, glycaemic control, BMI and high/treated blood pressure. The stroke and CKD models accounted for a moderate amount of the variation in outcomes observed (an R^2 of 0.35 and 0.34, respectively). The stroke model found that age, prior CHD, smoking and high/treated blood pressure were significant predictors of future stroke risk. The CKD model found that male gender, age, prior CHD and stroke were significant predictors of future CKD risk, as were smoking, glycaemic control, BMI and high/treated blood pressure. The CHD model had the smallest R^2 (0.09). Although it included known risk factors for CHD, the model accounted for little of the variation in outcomes between individuals and would not, therefore, be useful in clinical practice.

The cohort of patients used to populate the models appear to be representative of the wider UK population of people with type 2 diabetes, and unlike some previous models included

patients of all ages and health statuses. However incentives introduced as part of QOF since the end of the study period (1998-2003) may have improved the management of newly diagnosed diabetes in the years following diagnosis and may have resulted in improved outcomes in more recent years. It would therefore be prudent to update and extend these models using more recent clinical data and to assess their predictive validity in one or more external populations, particularly in comparison with existing risk models available in clinical practice.

APPENDICES

Appendix 2

Rapid review methods

Appendix 2.1 Methods and search terms used to identify previous CVD, CKD and all-cause mortality prediction models

Aim

To identify all papers presenting a CVD or CKD prediction model developed in patients with diabetes or that can be applied to individuals with type 2 diabetes.

CVD

1. Use van Dieren's existing systematic review.
2. Run van Dieren's search again in PubMed to identify any additional studies published in period 2011-2012.
3. Check PubMed suggested papers for papers identified in steps 1 & 2.

CKD and all-cause mortality

1. Run PubMed search to identify studies published in period 1991-2012.
2. Check PubMed suggested papers for papers identified in step 1.

Eligibility criteria

1. The prediction model was either developed in people with diabetes or included diabetes as a predictor.
2. The outcome of the prediction model was CVD or CKD or a CVD/CKD component (i.e., CHD, stroke, end stage renal failure, kidney dialysis, kidney transplant).
3. It presented a specific prediction rule/model with sufficient information on all variables to calculate the CVD CKD risk in a different population (beta coefficients of the model or otherwise a scoring system/graph/score card/nomogram was provided).

Exclusion criteria

1. Non-human studies.
2. Articles in languages other than English.
3. Studies presenting a prediction model developed in patients with previous CVD/CKD.
4. Studies focusing on the added predictive value of new risk factors to an existing prediction model.
5. Studies where full text was not available. These could not have presented sufficient information on all variables to calculate CKD risk.

Screening process

1. Screen on title.
2. Screen on abstract.
3. Screen on full text.

Search terms for CVD

((

Validat$ OR Predict$.ti. OR Rule$)

OR

(Predict$ AND (Outcome$ OR Risk$ OR Model$))

OR

(Decision$ AND (Model$ OR Clinical$ OR Logistic Models/))

OR

(Prognostic AND (History OR Variable$ OR Criteria OR Scor$ OR Characteristic$ OR Finding$ OR Factor$ OR Model$))

OR

("risk score"[All fields] OR "prediction model"[All fields] OR "prediction rule"[All fields] OR "risk assessment" [All fields] OR "algorithm"[All fields]

))

AND

(cardiovascular OR coronary OR cerebrovascular OR heart OR stroke)

AND

(diabetes OR "diabetes mellitus" OR "type 2 diabetes")

NOT

(Animals[MeSH] NOT Humans[MeSH])

Search terms for CKD

((

Validat$ OR Predict$.ti. OR Rule$)

OR

(Predict$ AND (Outcome$ OR Risk$ OR Model$))

OR

(Decision$ AND (Model$ OR Clinical$ OR Logistic Models/))

OR

(Prognostic AND (History OR Variable$ OR Criteria OR Scor$ OR Characteristic$ OR Finding$ OR Factor$ OR Model$))

OR

("risk score"[All fields] OR "prediction model"[All fields] OR "prediction rule"[All fields] OR "risk assessment" [All fields] OR "algorithm"[All fields]

))

AND

(CKD OR kidney OR nephr OR dialysis OR transplant OR replacement OR "end stage")

AND

(diabetes OR "diabetes mellitus" OR "type 2 diabetes")

NOT

(Animals[MeSH] NOT Humans[MeSH])

Search terms for all-cause mortality

((

Validat$ OR Predict$.ti. OR Rule$)

OR

(Predict$ AND (Outcome$ OR Risk$ OR Model$))

OR

(Decision$ AND (Model$ OR Clinical$ OR Logistic Models/))

OR

(Prognostic AND (History OR Variable$ OR Criteria OR Scor$ OR Characteristic$ OR Finding$ OR Factor$ OR Model$))

OR

("risk score"[All fields] OR "prediction model"[All fields] OR "prediction rule"[All fields] OR "risk assessment" [All fields] OR "algorithm"[All fields]

))

AND

(death OR mortality)

AND

(diabetes OR "diabetes mellitus" OR "type 2 diabetes")

NOT

(Animals[MeSH] NOT Humans[MeSH])

Appendix 6

Read codes used to identify cases of Type 2 diabetes mellitus

Note: This list includes codes specific to Type 2 diabetes, codes which do not specify diabetes type, and codes for Type 1 diabetes. These are used in combination with other criteria to identify cases of Type 2 diabetes: age at diagnosis >= 35 years and no insulin treatment within one year of diagnosis.

Read code	Description
13AB.00	Diabetic lipid lowering diet
13AC.00	Diabetic weight reducing diet
13B1.00	Diabetic diet
1434.00	H/O: diabetes mellitus
14F4.00	H/O: Admission in last year for diabetes foot problem
2BBF.00	Retinal abnormality - diabetes related
2BBL.00	O/E - diabetic maculopathy present both eyes
2BBP.00	O/E - right eye background diabetic retinopathy
2BBQ.00	O/E - left eye background diabetic retinopathy
2BBR.00	O/E - right eye preproliferative diabetic retinopathy
2BBS.00	O/E - left eye preproliferative diabetic retinopathy
2BBT.00	O/E - right eye proliferative diabetic retinopathy
2BBV.00	O/E - left eye proliferative diabetic retinopathy
2BBW.00	O/E - right eye diabetic maculopathy
2BBX.00	O/E - left eye diabetic maculopathy
2G51000	Foot abnormality - diabetes related
2G5A.00	O/E - Right diabetic foot at risk
2G5B.00	O/E - Left diabetic foot at risk

2G5C.00	Foot abnormality - diabetes related
2G5E.00	O/E - Right diabetic foot at low risk
2G5F.00	O/E - Right diabetic foot at moderate risk
2G5G.00	O/E - Right diabetic foot at high risk
2G5H.00	O/E - Right diabetic foot - ulcerated
2G5I.00	O/E - Left diabetic foot at low risk
2G5J.00	O/E - Left diabetic foot at moderate risk
2G5K.00	O/E - Left diabetic foot at high risk
2G5L.00	O/E - Left diabetic foot - ulcerated
3881.00	Education score - diabetes
3882.00	Diabetes well being questionnaire
3883.00	Diabetes treatment satisfaction questionnaire
42W..00	Hb. A1C - diabetic control
42W..11	Glycosylated Hb
42W..12	Glycated haemoglobin
42W1.00	Hb. A1C < 7% - good control
42W2.00	Hb. A1C 7-10% - borderline
42W3.00	Hb. A1C > 10% - bad control
42WZ.00	Hb. A1C - diabetic control NOS
42c..00	HbA1 - diabetic control
66A..00	Diabetic monitoring
66A1.00	Initial diabetic assessment
66A2.00	Follow-up diabetic assessment
66A3.00	Diabetic on diet only
66A4.00	Diabetic on oral treatment
66A5.00	Diabetic on insulin
66A8.00	Has seen dietician - diabetes
66A9.00	Understands diet - diabetes
66AA.11	Injection sites - diabetic

66AD.00	Fundoscopy - diabetic check
66AG.00	Diabetic drug side effects
66AH.00	Diabetic treatment changed
66AI.00	Diabetic - good control
66AJ.00	Diabetic - poor control
66AJ.11	Unstable diabetes
66AJ100	Brittle diabetes
66AJz00	Diabetic - poor control NOS
66AK.00	Diabetic - cooperative patient
66AL.00	Diabetic-uncooperative patient
66AM.00	Diabetic - follow-up default
66AN.00	Date diabetic treatment start
66AO.00	Date diabetic treatment stopp.
66AP.00	Diabetes: practice programme
66AQ.00	Diabetes: shared care programme
66AR.00	Diabetes management plan given
66AS.00	Diabetic annual review
66AT.00	Annual diabetic blood test
66AU.00	Diabetes care by hospital only
66AV.00	Diabetic on insulin and oral treatment
66AW.00	Diabetic foot risk assessment
66AX.00	Diabetes: shared care in pregnancy - diabetol and obstet
66AY.00	Diabetic diet - good compliance
66AZ.00	Diabetic monitoring NOS
66Aa.00	Diabetic diet - poor compliance
66Ab.00	Diabetic foot examination
66Ac.00	Diabetic peripheral neuropathy screening
8A12.00	Diabetic crisis monitoring
8A13.00	Diabetic stabilisation

8CA4100	Pt advised re diabetic diet
8H2J.00	Admit diabetic emergency
8H3O.00	Non-urgent diabetic admission
8H4F.00	Referral to diabetologist
8H7C.00	Refer, diabetic liaison nurse
8H7f.00	Referral to diabetes nurse
8HKE.00	Diabetology D.V. requested
8HLE.00	Diabetology D.V. done
8HME.00	Listed for Diabetology admissn
8HVU.00	Private referral to diabetologist
9N1v.00	Seen in diabetic eye clinic
9NM0.00	Attending diabetes clinic
9OL..00	Diabetes monitoring admin.
9OL..11	Diabetes clinic administration
9OL1.00	Attends diabetes monitoring
9OL2.00	Refuses diabetes monitoring
9OL3.00	Diabetes monitoring default
9OL4.00	Diabetes monitoring 1st letter
9OL5.00	Diabetes monitoring 2nd letter
9OL6.00	Diabetes monitoring 3rd letter
9OL7.00	Diabetes monitor.verbal invite
9OL8.00	Diabetes monitor.phone invite
9OL9.00	Diabetes monitoring deleted
9OLA.00	Diabetes monitor. check done
9OLA.11	Diabetes monitored
9OLZ.00	Diabetes monitoring admin.NOS
C10..00	Diabetes mellitus
C100.00	Diabetes mellitus with no mention of complication
C100000	Diabetes mellitus, juvenile type, no mention of complication

C100011	Insulin dependent diabetes mellitus
C100100	Diabetes mellitus, adult onset, no mention of complication
C100111	Maturity onset diabetes
C100112	Non-insulin dependent diabetes mellitus
C100z00	Diabetes mellitus NOS with no mention of complication
C101.00	Diabetes mellitus with ketoacidosis
C101000	Diabetes mellitus, juvenile type, with ketoacidosis
C101100	Diabetes mellitus, adult onset, with ketoacidosis
C101y00	Other specified diabetes mellitus with ketoacidosis
C101z00	Diabetes mellitus NOS with ketoacidosis
C102.00	Diabetes mellitus with hyperosmolar coma
C102000	Diabetes mellitus, juvenile type, with hyperosmolar coma
C102100	Diabetes mellitus, adult onset, with hyperosmolar coma
C102z00	Diabetes mellitus NOS with hyperosmolar coma
C103.00	Diabetes mellitus with ketoacidotic coma
C103000	Diabetes mellitus, juvenile type, with ketoacidotic coma
C103100	Diabetes mellitus, adult onset, with ketoacidotic coma
C103y00	Other specified diabetes mellitus with coma
C103z00	Diabetes mellitus NOS with ketoacidotic coma
C104.00	Diabetes mellitus with renal manifestation
C104.11	Diabetic nephropathy
C104000	Diabetes mellitus, juvenile type, with renal manifestation
C104100	Diabetes mellitus, adult onset, with renal manifestation
C104y00	Other specified diabetes mellitus with renal complications
C104z00	Diabetes mellitus with nephropathy NOS
C105.00	Diabetes mellitus with ophthalmic manifestation
C105000	Diabetes mellitus, juvenile type, + ophthalmic manifestation
C105100	Diabetes mellitus, adult onset, + ophthalmic manifestation
C105y00	Other specified diabetes mellitus with ophthalmic complicatn

C105z00	Diabetes mellitus NOS with ophthalmic manifestation
C106.00	Diabetes mellitus with neurological manifestation
C106.11	Diabetic amyotrophy
C106.12	Diabetes mellitus with neuropathy
C106.13	Diabetes mellitus with polyneuropathy
C106000	Diabetes mellitus, juvenile, + neurological manifestation
C106100	Diabetes mellitus, adult onset, + neurological manifestation
C106y00	Other specified diabetes mellitus with neurological comps
C106z00	Diabetes mellitus NOS with neurological manifestation
C107.00	Diabetes mellitus with peripheral circulatory disorder
C107.11	Diabetes mellitus with gangrene
C107.12	Diabetes with gangrene
C107000	Diabetes mellitus, juvenile +peripheral circulatory disorder
C107100	Diabetes mellitus, adult, + peripheral circulatory disorder
C107200	Diabetes mellitus, adult with gangrene
C107300	IDDM with peripheral circulatory disorder
C107400	NIDDM with peripheral circulatory disorder
C107y00	Other specified diabetes mellitus with periph circ comps
C107z00	Diabetes mellitus NOS with peripheral circulatory disorder
C108.00	Insulin dependent diabetes mellitus
C108.11	IDDM-Insulin dependent diabetes mellitus
C108.12	Type 1 diabetes mellitus
C108.13	Type I diabetes mellitus
C108000	Insulin-dependent diabetes mellitus with renal complications
C108011	Type I diabetes mellitus with renal complications
C108012	Type 1 diabetes mellitus with renal complications
C108100	Insulin-dependent diabetes mellitus with ophthalmic comps
C108111	Type I diabetes mellitus with ophthalmic complications
C108112	Type 1 diabetes mellitus with ophthalmic complications

C108200	Insulin-dependent diabetes mellitus with neurological comps
C108211	Type I diabetes mellitus with neurological complications
C108212	Type 1 diabetes mellitus with neurological complications
C108300	Insulin dependent diabetes mellitus with multiple complicatn
C108311	Type I diabetes mellitus with multiple complications
C108312	Type 1 diabetes mellitus with multiple complications
C108400	Unstable insulin dependent diabetes mellitus
C108411	Unstable type I diabetes mellitus
C108412	Unstable type 1 diabetes mellitus
C108500	Insulin dependent diabetes mellitus with ulcer
C108511	Type I diabetes mellitus with ulcer
C108512	Type 1 diabetes mellitus with ulcer
C108600	Insulin dependent diabetes mellitus with gangrene
C108611	Type I diabetes mellitus with gangrene
C108612	Type 1 diabetes mellitus with gangrene
C108700	Insulin dependent diabetes mellitus with retinopathy
C108711	Type I diabetes mellitus with retinopathy
C108712	Type 1 diabetes mellitus with retinopathy
C108800	Insulin dependent diabetes mellitus - poor control
C108811	Type I diabetes mellitus - poor control
C108812	Type 1 diabetes mellitus - poor control
C108900	Insulin dependent diabetes maturity onset
C108911	Type I diabetes mellitus maturity onset
C108912	Type 1 diabetes mellitus maturity onset
C108A00	Insulin-dependent diabetes without complication
C108A11	Type I diabetes mellitus without complication
C108A12	Type 1 diabetes mellitus without complication
C108B00	Insulin dependent diabetes mellitus with mononeuropathy
C108B11	Type I diabetes mellitus with mononeuropathy

C108B12	Type 1 diabetes mellitus with mononeuropathy
C108C00	Insulin dependent diabetes mellitus with polyneuropathy
C108C11	Type I diabetes mellitus with polyneuropathy
C108C12	Type 1 diabetes mellitus with polyneuropathy
C108D00	Insulin dependent diabetes mellitus with nephropathy
C108D11	Type I diabetes mellitus with nephropathy
C108D12	Type 1 diabetes mellitus with nephropathy
C108E00	Insulin dependent diabetes mellitus with hypoglycaemic coma
C108E11	Type I diabetes mellitus with hypoglycaemic coma
C108E12	Type 1 diabetes mellitus with hypoglycaemic coma
C108F00	Insulin dependent diabetes mellitus with diabetic cataract
C108F11	Type I diabetes mellitus with diabetic cataract
C108F12	Type 1 diabetes mellitus with diabetic cataract
C108G00	Insulin dependent diab mell with peripheral angiopathy
C108G11	Type I diabetes mellitus with peripheral angiopathy
C108G12	Type 1 diabetes mellitus with peripheral angiopathy
C108H00	Insulin dependent diabetes mellitus with arthropathy
C108H11	Type I diabetes mellitus with arthropathy
C108H12	Type 1 diabetes mellitus with arthropathy
C108J00	Insulin dependent diab mell with neuropathic arthropathy
C108J11	Type I diabetes mellitus with neuropathic arthropathy
C108J12	Type 1 diabetes mellitus with neuropathic arthropathy
C108y00	Other specified diabetes mellitus with multiple comps
C108z00	Unspecified diabetes mellitus with multiple complications
C109.00	Non-insulin dependent diabetes mellitus
C109.11	NIDDM - Non-insulin dependent diabetes mellitus
C109.12	Type 2 diabetes mellitus
C109.13	Type II diabetes mellitus
C109000	Non-insulin-dependent diabetes mellitus with renal comps

C109011	Type II diabetes mellitus with renal complications
C109012	Type 2 diabetes mellitus with renal complications
C109100	Non-insulin-dependent diabetes mellitus with ophthalm comps
C109111	Type II diabetes mellitus with ophthalmic complications
C109112	Type 2 diabetes mellitus with ophthalmic complications
C109200	Non-insulin-dependent diabetes mellitus with neuro comps
C109211	Type II diabetes mellitus with neurological complications
C109212	Type 2 diabetes mellitus with neurological complications
C109300	Non-insulin-dependent diabetes mellitus with multiple comps
C109311	Type II diabetes mellitus with multiple complications
C109312	Type 2 diabetes mellitus with multiple complications
C109400	Non-insulin dependent diabetes mellitus with ulcer
C109411	Type II diabetes mellitus with ulcer
C109412	Type 2 diabetes mellitus with ulcer
C109500	Non-insulin dependent diabetes mellitus with gangrene
C109511	Type II diabetes mellitus with gangrene
C109512	Type 2 diabetes mellitus with gangrene
C109600	Non-insulin-dependent diabetes mellitus with retinopathy
C109611	Type II diabetes mellitus with retinopathy
C109612	Type 2 diabetes mellitus with retinopathy
C109700	Non-insulin dependent diabetes mellitus - poor control
C109711	Type II diabetes mellitus - poor control
C109712	Type 2 diabetes mellitus - poor control
C109800	Reaven's syndrome
C109900	Non-insulin-dependent diabetes mellitus without complication
C109911	Type II diabetes mellitus without complication
C109912	Type 2 diabetes mellitus without complication
C109A00	Non-insulin dependent diabetes mellitus with mononeuropathy
C109A11	Type II diabetes mellitus with mononeuropathy

C109A12	Type 2 diabetes mellitus with mononeuropathy
C109B00	Non-insulin dependent diabetes mellitus with polyneuropathy
C109B11	Type II diabetes mellitus with polyneuropathy
C109B12	Type 2 diabetes mellitus with polyneuropathy
C109C00	Non-insulin dependent diabetes mellitus with nephropathy
C109C11	Type II diabetes mellitus with nephropathy
C109C12	Type 2 diabetes mellitus with nephropathy
C109D00	Non-insulin dependent diabetes mellitus with hypoglyca coma
C109D11	Type II diabetes mellitus with hypoglycaemic coma
C109D12	Type 2 diabetes mellitus with hypoglycaemic coma
C109E00	Non-insulin depend diabetes mellitus with diabetic cataract
C109E11	Type II diabetes mellitus with diabetic cataract
C109E12	Type 2 diabetes mellitus with diabetic cataract
C109F00	Non-insulin-dependent d m with peripheral angiopath
C109F11	Type II diabetes mellitus with peripheral angiopathy
C109F12	Type 2 diabetes mellitus with peripheral angiopathy
C109G00	Non-insulin dependent diabetes mellitus with arthropathy
C109G11	Type II diabetes mellitus with arthropathy
C109G12	Type 2 diabetes mellitus with arthropathy
C109H00	Non-insulin dependent d m with neuropathic arthropathy
C109H11	Type II diabetes mellitus with neuropathic arthropathy
C109H12	Type 2 diabetes mellitus with neuropathic arthropathy
C109J00	Insulin treated Type 2 diabetes mellitus
C109J11	Insulin treated non-insulin dependent diabetes mellitus
C109J12	Insulin treated Type II diabetes mellitus
C109K00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus
C10E.00	Type 1 diabetes mellitus
C10E.11	Type I diabetes mellitus
C10E.12	Insulin dependent diabetes mellitus

C10E000	Type 1 diabetes mellitus with renal complications
C10E011	Type I diabetes mellitus with renal complications
C10E012	Insulin-dependent diabetes mellitus with renal complications
C10E100	Type 1 diabetes mellitus with ophthalmic complications
C10E111	Type I diabetes mellitus with ophthalmic complications
C10E112	Insulin-dependent diabetes mellitus with ophthalmic comps
C10E200	Type 1 diabetes mellitus with neurological complications
C10E211	Type I diabetes mellitus with neurological complications
C10E212	Insulin-dependent diabetes mellitus with neurological comps
C10E400	Unstable type 1 diabetes mellitus
C10E411	Unstable type I diabetes mellitus
C10E412	Unstable insulin dependent diabetes mellitus
C10E500	Type 1 diabetes mellitus with ulcer
C10E511	Type I diabetes mellitus with ulcer
C10E512	Insulin dependent diabetes mellitus with ulcer
C10E600	Type 1 diabetes mellitus with gangrene
C10E611	Type I diabetes mellitus with gangrene
C10E612	Insulin dependent diabetes mellitus with gangrene
C10E700	Type 1 diabetes mellitus with retinopathy
C10E711	Type I diabetes mellitus with retinopathy
C10E712	Insulin dependent diabetes mellitus with retinopathy
C10E800	Type 1 diabetes mellitus - poor control
C10E811	Type I diabetes mellitus - poor control
C10E812	Insulin dependent diabetes mellitus - poor control
C10E900	Type 1 diabetes mellitus maturity onset
C10E911	Type I diabetes mellitus maturity onset
C10E912	Insulin dependent diabetes maturity onset
C10EC00	Type 1 diabetes mellitus with polyneuropathy
C10EC11	Type I diabetes mellitus with polyneuropathy

C10EC12	Insulin dependent diabetes mellitus with polyneuropathy
C10ED00	Type 1 diabetes mellitus with nephropathy
C10ED11	Type I diabetes mellitus with nephropathy
C10ED12	Insulin dependent diabetes mellitus with nephropathy
C10EE00	Type 1 diabetes mellitus with hypoglycaemic coma
C10EE11	Type I diabetes mellitus with hypoglycaemic coma
C10EE12	Insulin dependent diabetes mellitus with hypoglycaemic coma
C10EF00	Type 1 diabetes mellitus with diabetic cataract
C10EF11	Type I diabetes mellitus with diabetic cataract
C10EF12	Insulin dependent diabetes mellitus with diabetic cataract
C10EH00	Type 1 diabetes mellitus with arthropathy
C10EH11	Type I diabetes mellitus with arthropathy
C10EH12	Insulin dependent diabetes mellitus with arthropathy
C10EJ00	Type 1 diabetes mellitus with neuropathic arthropathy
C10EJ11	Type I diabetes mellitus with neuropathic arthropathy
C10EJ12	Insulin dependent diab mell with neuropathic arthropathy
C10EK00	Type 1 diabetes mellitus with persistent proteinuria
C10EK11	Type I diabetes mellitus with persistent proteinuria
C10EL00	Type 1 diabetes mellitus with persistent microalbuminuria
C10EL11	Type I diabetes mellitus with persistent microalbuminuria
C10EM00	Type 1 diabetes mellitus with ketoacidosis
C10EM11	Type I diabetes mellitus with ketoacidosis
C10EN00	Type 1 diabetes mellitus with ketoacidotic coma
C10EN11	Type I diabetes mellitus with ketoacidotic coma
C10EP00	Type 1 diabetes mellitus with exudative maculopathy
C10EP11	Type I diabetes mellitus with exudative maculopathy
C10F.00	Type 2 diabetes mellitus
C10F.11	Type II diabetes mellitus
C10F000	Type 2 diabetes mellitus with renal complications

C10F011	Type II diabetes mellitus with renal complications
C10F100	Type 2 diabetes mellitus with ophthalmic complications
C10F111	Type II diabetes mellitus with ophthalmic complications
C10F200	Type 2 diabetes mellitus with neurological complications
C10F211	Type II diabetes mellitus with neurological complications
C10F300	Type 2 diabetes mellitus with multiple complications
C10F311	Type II diabetes mellitus with multiple complications
C10F400	Type 2 diabetes mellitus with ulcer
C10F411	Type II diabetes mellitus with ulcer
C10F500	Type 2 diabetes mellitus with gangrene
C10F511	Type II diabetes mellitus with gangrene
C10F600	Type 2 diabetes mellitus with retinopathy
C10F611	Type II diabetes mellitus with retinopathy
C10F700	Type 2 diabetes mellitus - poor control
C10F711	Type II diabetes mellitus - poor control
C10F800	Reaven's syndrome
C10F811	Metabolic syndrome X
C10F900	Type 2 diabetes mellitus without complication
C10F911	Type II diabetes mellitus without complication
C10FA00	Type 2 diabetes mellitus with mononeuropathy
C10FA11	Type II diabetes mellitus with mononeuropathy
C10FB00	Type 2 diabetes mellitus with polyneuropathy
C10FB11	Type II diabetes mellitus with polyneuropathy
C10FC00	Type 2 diabetes mellitus with nephropathy
C10FC11	Type II diabetes mellitus with nephropathy
C10FD00	Type 2 diabetes mellitus with hypoglycaemic coma
C10FD11	Type II diabetes mellitus with hypoglycaemic coma
C10FE00	Type 2 diabetes mellitus with diabetic cataract
C10FE11	Type II diabetes mellitus with diabetic cataract

C10FF00	Type 2 diabetes mellitus with peripheral angiopathy
C10FF11	Type II diabetes mellitus with peripheral angiopathy
C10FG00	Type 2 diabetes mellitus with arthropathy
C10FG11	Type II diabetes mellitus with arthropathy
C10FH00	Type 2 diabetes mellitus with neuropathic arthropathy
C10FH11	Type II diabetes mellitus with neuropathic arthropathy
C10FJ00	Insulin treated Type 2 diabetes mellitus
C10FJ11	Insulin treated Type II diabetes mellitus
C10FK00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus
C10FL00	Type 2 diabetes mellitus with persistent proteinuria
C10FL11	Type II diabetes mellitus with persistent proteinuria
C10FM00	Type 2 diabetes mellitus with persistent microalbuminuria
C10FM11	Type II diabetes mellitus with persistent microalbuminuria
C10FN00	Type 2 diabetes mellitus with ketoacidosis
C10FN11	Type II diabetes mellitus with ketoacidosis
C10FP00	Type 2 diabetes mellitus with ketoacidotic coma
C10FP11	Type II diabetes mellitus with ketoacidotic coma
C10FQ00	Type 2 diabetes mellitus with exudative maculopathy
C10FQ11	Type II diabetes mellitus with exudative maculopathy
C10G.00	Secondary pancreatic diabetes mellitus
C10K.00	Type A insulin resistance
C10M.00	Lipoatrophic diabetes mellitus
C10y.00	Diabetes mellitus with other specified manifestation
C10y000	Diabetes mellitus, juvenile, + other specified manifestation
C10y100	Diabetes mellitus, adult, + other specified manifestation
C10yy00	Other specified diabetes mellitus with other spec comps
C10yz00	Diabetes mellitus NOS with other specified manifestation
C10z.00	Diabetes mellitus with unspecified complication
C10z000	Diabetes mellitus, juvenile type, + unspecified complication

C10z100	Diabetes mellitus, adult onset, + unspecified complication
C10zy00	Other specified diabetes mellitus with unspecified comps
C10zz00	Diabetes mellitus NOS with unspecified complication
Cyu2.00	[X]Diabetes mellitus
Cyu2300	[X]Unspecified diabetes mellitus with renal complications
F171100	Autonomic neuropathy due to diabetes
F345000	Diabetic mononeuritis multiplex
F35z000	Diabetic mononeuritis NOS
F372.00	Polyneuropathy in diabetes
F372.11	Diabetic polyneuropathy
F372.12	Diabetic neuropathy
F372000	Acute painful diabetic neuropathy
F372100	Chronic painful diabetic neuropathy
F372200	Asymptomatic diabetic neuropathy
F381300	Myasthenic syndrome due to diabetic amyotrophy
F381311	Diabetic amyotrophy
F3y0.00	Diabetic mononeuropathy
F420.00	Diabetic retinopathy
F420000	Background diabetic retinopathy
F420100	Proliferative diabetic retinopathy
F420200	Preproliferative diabetic retinopathy
F420300	Advanced diabetic maculopathy
F420400	Diabetic maculopathy
F420500	Advanced diabetic retinal disease
F420600	Non proliferative diabetic retinopathy
F420700	High risk proliferative diabetic retinopathy
F420800	High risk non proliferative diabetic retinopathy
F420z00	Diabetic retinopathy NOS
F440700	Diabetic iritis

F464000	Diabetic cataract
G73y000	Diabetic peripheral angiopathy
K01x100	Nephrotic syndrome in diabetes mellitus
K01x111	Kimmelstiel - Wilson disease
Kyu0300	[X]Glomerular disorders in diabetes mellitus
L180500	Pre-existing diabetes mellitus, insulin-dependent
L180600	Pre-existing diabetes mellitus, non-insulin-dependent
L180700	Pre-existing malnutrition-related diabetes mellitus
L180X00	Pre-existing diabetes mellitus, unspecified
Lyu2900	[X]Pre-existing diabetes mellitus, unspecified
M037200	Cellulitis in diabetic foot
M271000	Ischaemic ulcer diabetic foot
M271100	Neuropathic diabetic ulcer - foot
M271200	Mixed diabetic ulcer - foot
N030000	Diabetic cheiroarthropathy
N030011	Diabetic cheiropathy
N030100	Diabetic Charcot arthropathy
R054200	[D]Gangrene of toe in diabetic
R054300	[D]Widespread diabetic foot gangrene
ZV65312	[V]Dietary counselling in diabetes mellitus

Read codes used to identify pregnancy

Note: First 100 codes from list of 3154 presented here to demonstrate range of codes in full table.

Read code	Description
13H7.00	Unwanted pregnancy
13H7.11	Unwanted child
13H8.00	Illegitimate pregnancy
13S..00	Pregnancy benefits
13S..11	Maternity allowances
13SZ.00	Pregnancy benefit NOS
1514.00	Estimated date of confinement
1514.11	Due to deliver - EDC
1514.12	Estimated date of delivery
27...00	Obstetric examination
271..00	O/E - gravid uterus size
271..11	O/E - fundus size - obstetric
271..12	O/E - uterus size - obstetric
272..00	O/E - fetal presentation
272..11	O/E - lie of fetus
272..12	O/E - presenting part
2726.00	O/E -fetal presentation unsure
272Z.00	O/E - fetal presentation NOS
274Z.00	O/E - fetal station NOS
275..00	O/E - fetal movements
2751.00	O/E - no fetal movements
2752.00	O/E - fetal movements seen

2753.00	O/E - fetal movements felt
2754.00	O/E - fetus very active
2755.00	O/E - fetal movemnt.diminished
275Z.00	O/E - fetal movements NOS
276..00	O/E - fetal heart heard
2761.00	O/E - fetal heart not heard
2762.00	O/E - fetal heart < 40
2763.00	O/E - fetal heart 40-80
2764.00	O/E - fetal heart 80-100
2765.00	O/E - fetal heart 100-120
2766.00	O/E - fetal heart 120-160
2767.00	O/E - fetal heart 160-180
2768.00	O/E - fetal heart 180-200
2769.00	O/E - fetal heart > 200
276A.00	O/E - fetal heart -type 1 dips
276B.00	O/E - fetal heart -type 2 dips
276Z.00	O/E - fetal heart NOS
27A..00	O/E - VE - descent of P. part
27A..11	O/E - VE - descent of fetus
27A..12	O/E - VE - presenting part
27B..00	O/E - viable fetus
27Z..00	Obstetric examination NOS
3188.00	Placental localisation
3885.00	Edinburgh postnatal depression scale
444..00	Feto/placental hormones
4441.00	Feto/placent. hormones abnorm.
4442.00	Feto/placen. hormones normal
4443.00	Placental lactogen - HPL
4443.11	HPL - Human placental lactogen level

4443000	Human placental lactogen level normal
4443100	HPL - Human placental lactogen abnormal
4444.00	Serum oestriol level
4444.11	Human placental lactogen
4444.12	Placental lactogen
4445.00	Placental function test
4445000	Placental function test normal
4445100	Placental function test abnormal
444Z.00	Feto/placental hormones NOS
4453.00	Serum pregnancy test positive
4654.00	Urine pregnancy test positive
4H...00	Amniotic fluid examination
4H...11	Liquor examination
4H1..00	Amniotic fluid exam. - general
4H11.00	Amniotic fluid sent for exam.
4H12.00	Amniotic fluid - nil abnormal
4H13.00	Amniotic fluid - abnormality
4H1Z.00	Amniotic fluid exam. gen. NOS
4H2..00	Amniotic fluid appearance
4H21.00	Amniotic fluid - clear
4H22.00	Amniotic fluid - blood stained
4H23.00	Amniotic fluid -meconium stain
4H2Z.00	Amniotic fluid appearance NOS
4H3..00	Amniotic fluid microscopy
4H31.00	Amniotic fluid microscopy -NAD
4H32.00	Amniotic fluid microsc. - abn.
4H33.00	Amniotic fluid cell content OK
4H3Z.00	Amniotic fluid microscopy NOS
4H4..00	Amniotic fluid chemistry

4H41.00	Amniotic fluid chemistry: NAD
4H42.00	Amniotic fluid chemistry: abn.
4H43.00	Amniotic fluid L/S ratio
4H43.11	Lecithin - amniotic
4H43.12	Sphingomyelin -amniotic
4H44.00	Amniotic fluid palmitic acid
4H45.00	Amniotic fluid cholinesterase
4H4Z.00	Amniotic fluid chemistry NOS
4H5..00	Amniotic fluid AFP
4H51.00	Amniotic fluid AFP normal
4H52.00	Amniotic fluid AFP equivocal
4H53.00	Amniotic fluid AFP abnormal
4H5Z.00	Amniotic fluid AFP NOS
4H7..00	Amniotic fetal cell study
4H71.00	Amniotic fetal cell study: NAD
4H72.00	Amniotic fetal cell abnormal
4H73.00	Amniotic fetal cell: mongol
4H7Z.00	Amniotic fetal cell study NOS
4HZ..00	Amniotic fluid exam. NOS
4JL3.00	Amniotic fluid for organism

Read codes used to identify cases of CHD

Read code	Description
14A3.00	H/O: myocardial infarct <60
14A4.00	H/O: myocardial infarct >60
14A5.00	H/O: angina pectoris
14AH.00	H/O: Myocardial infarction in last year
14AJ.00	H/O: Angina in last year
14AL.00	H/O: Treatment for ischaemic heart disease
322..00	ECG: myocardial ischaemia
3222.00	ECG:shows myocardial ischaemia
322Z.00	ECG: myocardial ischaemia NOS
323..00	ECG: myocardial infarction
3232.00	ECG: old myocardial infarction
3233.00	ECG: antero-septal infarct.
3234.00	ECG:posterior/inferior infarct
3235.00	ECG: subendocardial infarct
3236.00	ECG: lateral infarction
323Z.00	ECG: myocardial infarct NOS
44H3.00	Cardiac enzymes abnormal
44H3000	Cardiac enzymes abnormal - first set
5543.00	Coronary arteriograph.abnormal
662K.00	Angina control
662K000	Angina control - good
662K100	Angina control - poor
662K200	Angina control - improving
662K300	Angina control - worsening
662Kz00	Angina control NOS

790H300	Revascularisation of wall of heart
792..00	Coronary artery operations
792..11	Coronary artery bypass graft operations
7920.00	Saphenous vein graft replacement of coronary artery
7920.11	Saphenous vein graft bypass of coronary artery
7920000	Saphenous vein graft replacement of one coronary artery
7920100	Saphenous vein graft replacement of two coronary arteries
7920200	Saphenous vein graft replacement of three coronary arteries
7920300	Saphenous vein graft replacement of four+ coronary arteries
7920y00	Saphenous vein graft replacement of coronary artery OS
7920z00	Saphenous vein graft replacement coronary artery NOS
7921.00	Other autograft replacement of coronary artery
7921.11	Other autograft bypass of coronary artery
7921000	Autograft replacement of one coronary artery NEC
7921100	Autograft replacement of two coronary arteries NEC
7921200	Autograft replacement of three coronary arteries NEC
7921300	Autograft replacement of four of more coronary arteries NEC
7921y00	Other autograft replacement of coronary artery OS
7921z00	Other autograft replacement of coronary artery NOS
7922.00	Allograft replacement of coronary artery
7922.11	Allograft bypass of coronary artery
7922000	Allograft replacement of one coronary artery
7922100	Allograft replacement of two coronary arteries
7922200	Allograft replacement of three coronary arteries
7922300	Allograft replacement of four or more coronary arteries
7922y00	Other specified allograft replacement of coronary artery
7922z00	Allograft replacement of coronary artery NOS
7923.00	Prosthetic replacement of coronary artery
7923.11	Prosthetic bypass of coronary artery

7923000	Prosthetic replacement of one coronary artery
7923100	Prosthetic replacement of two coronary arteries
7923200	Prosthetic replacement of three coronary arteries
7923300	Prosthetic replacement of four or more coronary arteries
7923y00	Other specified prosthetic replacement of coronary artery
7923z00	Prosthetic replacement of coronary artery NOS
7924.00	Revision of bypass for coronary artery
7924000	Revision of bypass for one coronary artery
7924100	Revision of bypass for two coronary arteries
7924200	Revision of bypass for three coronary arteries
7924300	Revision of bypass for four or more coronary arteries
7924400	Revision of connection of thoracic artery to coronary artery
7924500	Revision of implantation of thoracic artery into heart
7924y00	Other specified revision of bypass for coronary artery
7924z00	Revision of bypass for coronary artery NOS
7925.00	Connection of mammary artery to coronary artery
7925.11	Creation of bypass from mammary artery to coronary artery
7925000	Double anastomosis of mammary arteries to coronary arteries
7925011	LIMA sequential anastomosis
7925012	RIMA sequential anastomosis
7925100	Double implant of mammary arteries into coronary arteries
7925200	Single anast mammary art to left ant descend coronary art
7925300	Single anastomosis of mammary artery to coronary artery NEC
7925311	LIMA single anastomosis
7925312	RIMA single anastomosis
7925400	Single implantation of mammary artery into coronary artery
7925y00	Connection of mammary artery to coronary artery OS
7925z00	Connection of mammary artery to coronary artery NOS
7926.00	Connection of other thoracic artery to coronary artery

7926000	Double anastom thoracic arteries to coronary arteries NEC
7926100	Double implant thoracic arteries into coronary arteries NEC
7926200	Single anastomosis of thoracic artery to coronary artery NEC
7926300	Single implantation thoracic artery into coronary artery NEC
7926y00	Connection of other thoracic artery to coronary artery OS
7926z00	Connection of other thoracic artery to coronary artery NOS
7927.00	Other open operations on coronary artery
7927500	Open angioplasty of coronary artery
7928.00	Transluminal balloon angioplasty of coronary artery
7928.11	Percutaneous balloon coronary angioplasty
7928000	Percut transluminal balloon angioplasty one coronary artery
7928100	Percut translum balloon angioplasty mult coronary arteries
7928200	Percut translum balloon angioplasty bypass graft coronary a
7928y00	Transluminal balloon angioplasty of coronary artery OS
7928z00	Transluminal balloon angioplasty of coronary artery NOS
7929.00	Other therapeutic transluminal operations on coronary artery
7929000	Percutaneous transluminal laser coronary angioplasty
7929100	Percut transluminal coronary thrombolysis with streptokinase
7929111	Percut translum coronary thrombolytic therapy- streptokinase
7929200	Percut translum inject therap subst to coronary artery NEC
7929300	Rotary blade coronary angioplasty
7929400	Insertion of coronary artery stent
7929y00	Other therapeutic transluminal op on coronary artery OS
7929z00	Other therapeutic transluminal op on coronary artery NOS
792B.00	Repair of coronary artery NEC
792B000	Endarterectomy of coronary artery NEC
792By00	Other specified repair of coronary artery
792Bz00	Repair of coronary artery NOS
792C.00	Other replacement of coronary artery

792C000	Replacement of coronary arteries using multiple methods
792Cy00	Other specified replacement of coronary artery
792Cz00	Replacement of coronary artery NOS
792D.00	Other bypass of coronary artery
792Dy00	Other specified other bypass of coronary artery
792Dz00	Other bypass of coronary artery NOS
792y.00	Other specified operations on coronary artery
792z.00	Coronary artery operations NOS
88A8.00	Thrombolytic therapy
88A8.11	Fibrinolysis
8B27.00	Antianginal therapy
8B3k.00	Coronary heart disease medication review
8B63.11	Aspirin prophylaxis - IHD
G3...00	Ischaemic heart disease
G3...11	Arteriosclerotic heart disease
G3...12	Atherosclerotic heart disease
G3...13	IHD - Ischaemic heart disease
G30..00	Acute myocardial infarction
G30..11	Attack - heart
G30..12	Coronary thrombosis
G30..13	Cardiac rupture following myocardial infarction (MI)
G30..14	Heart attack
G30..15	MI - acute myocardial infarction
G30..16	Thrombosis - coronary
G30..17	Silent myocardial infarction
G300.00	Acute anterolateral infarction
G301.00	Other specified anterior myocardial infarction
G301000	Acute anteroapical infarction
G301100	Acute anteroseptal infarction

G301z00	Anterior myocardial infarction NOS
G302.00	Acute inferolateral infarction
G303.00	Acute inferoposterior infarction
G304.00	Posterior myocardial infarction NOS
G305.00	Lateral myocardial infarction NOS
G306.00	True posterior myocardial infarction
G307.00	Acute subendocardial infarction
G307000	Acute non-Q wave infarction
G307100	Acute non-ST segment elevation myocardial infarction
G308.00	Inferior myocardial infarction NOS
G309.00	Acute Q-wave infarct
G30A.00	Mural thrombosis
G30B.00	Acute posterolateral myocardial infarction
G30X.00	Acute transmural myocardial infarction of unspecif site
G30X000	Acute ST segment elevation myocardial infarction
G30y.00	Other acute myocardial infarction
G30y000	Acute atrial infarction
G30y100	Acute papillary muscle infarction
G30y200	Acute septal infarction
G30yz00	Other acute myocardial infarction NOS
G30z.00	Acute myocardial infarction NOS
G31..00	Other acute and subacute ischaemic heart disease
G310.00	Postmyocardial infarction syndrome
G310.11	Dressler's syndrome
G311.00	Preinfarction syndrome
G311.11	Crescendo angina
G311.12	Impending infarction
G311.13	Unstable angina
G311.14	Angina at rest

G311000	Myocardial infarction aborted
G311011	MI - myocardial infarction aborted
G311100	Unstable angina
G311200	Angina at rest
G311300	Refractory angina
G311400	Worsening angina
G311500	Acute coronary syndrome
G311z00	Preinfarction syndrome NOS
G312.00	Coronary thrombosis not resulting in myocardial infarction
G31y.00	Other acute and subacute ischaemic heart disease
G31y000	Acute coronary insufficiency
G31y100	Microinfarction of heart
G31y200	Subendocardial ischaemia
G31y300	Transient myocardial ischaemia
G31yz00	Other acute and subacute ischaemic heart disease NOS
G32..00	Old myocardial infarction
G32..11	Healed myocardial infarction
G32..12	Personal history of myocardial infarction
G33..00	Angina pectoris
G330.00	Angina decubitus
G330000	Nocturnal angina
G330z00	Angina decubitus NOS
G331.00	Prinzmetal's angina
G331.11	Variant angina pectoris
G332.00	Coronary artery spasm
G33z.00	Angina pectoris NOS
G33z000	Status anginosus
G33z100	Stenocardia
G33z200	Syncope anginosa

G33z300	Angina on effort
G33z400	Ischaemic chest pain
G33z500	Post infarct angina
G33z600	New onset angina
G33z700	Stable angina
G33zz00	Angina pectoris NOS
G34..00	Other chronic ischaemic heart disease
G340.00	Coronary atherosclerosis
G340.11	Triple vessel disease of the heart
G340.12	Coronary artery disease
G340000	Single coronary vessel disease
G340100	Double coronary vessel disease
G341.00	Aneurysm of heart
G341.11	Cardiac aneurysm
G341000	Ventricular cardiac aneurysm
G341100	Other cardiac wall aneurysm
G341111	Mural cardiac aneurysm
G341200	Aneurysm of coronary vessels
G341300	Acquired atrioventricular fistula of heart
G341z00	Aneurysm of heart NOS
G342.00	Atherosclerotic cardiovascular disease
G343.00	Ischaemic cardiomyopathy
G344.00	Silent myocardial ischaemia
G34y.00	Other specified chronic ischaemic heart disease
G34y000	Chronic coronary insufficiency
G34y100	Chronic myocardial ischaemia
G34yz00	Other specified chronic ischaemic heart disease NOS
G34z.00	Other chronic ischaemic heart disease NOS
G34z000	Asymptomatic coronary heart disease

G35..00	Subsequent myocardial infarction
G350.00	Subsequent myocardial infarction of anterior wall
G351.00	Subsequent myocardial infarction of inferior wall
G353.00	Subsequent myocardial infarction of other sites
G35X.00	Subsequent myocardial infarction of unspecified site
G36..00	Certain current complication follow acute myocardial infarct
G360.00	Haemopericardium/current comp folow acut myocard infarct
G361.00	Atrial septal defect/curr comp folow acut myocardal infarct
G362.00	Ventric septal defect/curr comp fol acut myocardal infarctn
G363.00	Ruptur cardiac wall w'out haemopericard/cur comp fol ac MI
G364.00	Ruptur chordae tendinae/curr comp fol acute myocard infarct
G365.00	Rupture papillary muscle/curr comp fol acute myocard infarct
G366.00	Thrombosis atrium,auric append&vent/curr comp foll acute MI
G37..00	Cardiac syndrome X
G38..00	Postoperative myocardial infarction
G380.00	Postoperative transmural myocardial infarction anterior wall
G381.00	Postoperative transmural myocardial infarction inferior wall
G382.00	Postoperative transmural myocardial infarction other sites
G383.00	Postoperative transmural myocardial infarction unspec site
G384.00	Postoperative subendocardial myocardial infarction
G38z.00	Postoperative myocardial infarction, unspecified
G3y..00	Other specified ischaemic heart disease
G3z..00	Ischaemic heart disease NOS
Gyu3.00	[X]Ischaemic heart diseases
Gyu3000	[X]Other forms of angina pectoris
Gyu3100	[X]Other current complicatns following acute myocard infarct
Gyu3200	[X]Other forms of acute ischaemic heart disease
Gyu3300	[X]Other forms of chronic ischaemic heart disease
Gyu3400	[X]Acute transmural myocardial infarction of unspecif site

Gyu3500	[X]Subsequent myocardial infarction of other sites
Gyu3600	[X]Subsequent myocardial infarction of unspecified site
SP00300	Mechanical complication of coronary bypass

Read codes used to identify cases of stroke

Read code	Description
G6...00	Cerebrovascular disease
G60..00	Subarachnoid haemorrhage
G600.00	Ruptured berry aneurysm
G601.00	Subarachnoid haemorrhage from carotid siphon and bifurcation
G602.00	Subarachnoid haemorrhage from middle cerebral artery
G603.00	Subarachnoid haemorrhage from anterior communicating artery
G604.00	Subarachnoid haemorrhage from posterior communicating artery
G605.00	Subarachnoid haemorrhage from basilar artery
G606.00	Subarachnoid haemorrhage from vertebral artery
G60X.00	Subarachnoid haemorrh from intracranial artery, unspecif
G60z.00	Subarachnoid haemorrhage NOS
G61..00	Intracerebral haemorrhage
G61..11	CVA - cerebrovascular accid due to intracerebral haemorrhage
G61..12	Stroke due to intracerebral haemorrhage
G610.00	Cortical haemorrhage
G611.00	Internal capsule haemorrhage
G612.00	Basal nucleus haemorrhage
G613.00	Cerebellar haemorrhage
G614.00	Pontine haemorrhage
G615.00	Bulbar haemorrhage
G616.00	External capsule haemorrhage
G617.00	Intracerebral haemorrhage, intraventricular
G618.00	Intracerebral haemorrhage, multiple localized
G61X.00	Intracerebral haemorrhage in hemisphere, unspecified
G61X000	Left sided intracerebral haemorrhage, unspecified

G61X100	Right sided intracerebral haemorrhage, unspecified
G61z.00	Intracerebral haemorrhage NOS
G62..00	Other and unspecified intracranial haemorrhage
G620.00	Extradural haemorrhage - nontraumatic
G621.00	Subdural haemorrhage - nontraumatic
G622.00	Subdural haematoma - nontraumatic
G623.00	Subdural haemorrhage NOS
G62z.00	Intracranial haemorrhage NOS
G63..00	Precerebral arterial occlusion
G63..11	Infarction - precerebral
G630.00	Basilar artery occlusion
G631.00	Carotid artery occlusion
G631.12	Thrombosis, carotid artery
G632.00	Vertebral artery occlusion
G633.00	Multiple and bilateral precerebral arterial occlusion
G63y.00	Other precerebral artery occlusion
G63y000	Cerebral infarct due to thrombosis of precerebral arteries
G63y100	Cerebral infarction due to embolism of precerebral arteries
G63z.00	Precerebral artery occlusion NOS
G64..00	Cerebral arterial occlusion
G64..11	CVA - cerebral artery occlusion
G64..12	Infarction - cerebral
G64..13	Stroke due to cerebral arterial occlusion
G640.00	Cerebral thrombosis
G640000	Cerebral infarction due to thrombosis of cerebral arteries
G641.00	Cerebral embolism
G641.11	Cerebral embolus
G641000	Cerebral infarction due to embolism of cerebral arteries
G64z.00	Cerebral infarction NOS

G64z.11	Brainstem infarction NOS
G64z.12	Cerebellar infarction
G64z000	Brainstem infarction
G64z100	Wallenberg syndrome
G64z111	Lateral medullary syndrome
G64z200	Left sided cerebral infarction
G64z300	Right sided cerebral infarction
G64z400	Infarction of basal ganglia
G66..00	Stroke and cerebrovascular accident unspecified
G66..11	CVA unspecified
G66..12	Stroke unspecified
G66..13	CVA - Cerebrovascular accident unspecified
G660.00	Middle cerebral artery syndrome
G661.00	Anterior cerebral artery syndrome
G662.00	Posterior cerebral artery syndrome
G663.00	Brain stem stroke syndrome
G664.00	Cerebellar stroke syndrome
G665.00	Pure motor lacunar syndrome
G666.00	Pure sensory lacunar syndrome
G667.00	Left sided CVA
G668.00	Right sided CVA
G669.00	Cerebral palsy, not congenital or infantile, acute
G67..00	Other cerebrovascular disease
G670.00	Cerebral atherosclerosis
G670.11	Precerebral atherosclerosis
G671.00	Generalised ischaemic cerebrovascular disease NOS
G671000	Acute cerebrovascular insufficiency NOS
G671100	Chronic cerebral ischaemia
G671z00	Generalised ischaemic cerebrovascular disease NOS

G672.00	Hypertensive encephalopathy
G673.00	Cerebral aneurysm, nonruptured
G673000	Dissection of cerebral arteries, nonruptured
G673100	Carotico-cavernous sinus fistula
G674.00	Cerebral arteritis
G674000	Cerebral amyloid angiopathy
G675.00	Moyamoya disease
G676.00	Nonpyogenic venous sinus thrombosis
G676000	Cereb infarct due cerebral venous thrombosis, nonpyogenic
G677.00	Occlusion/stenosis cerebral arts not result cerebral infarct
G677000	Occlusion and stenosis of middle cerebral artery
G677100	Occlusion and stenosis of anterior cerebral artery
G677200	Occlusion and stenosis of posterior cerebral artery
G677300	Occlusion and stenosis of cerebellar arteries
G677400	Occlusion+stenosis of multiple and bilat cerebral arteries
G678.00	Cereb autosom dominant arteriop subcort infarcts leukoenceph
G67y.00	Other cerebrovascular disease OS
G67z.00	Other cerebrovascular disease NOS
G68..00	Late effects of cerebrovascular disease
G680.00	Sequelae of subarachnoid haemorrhage
G681.00	Sequelae of intracerebral haemorrhage
G682.00	Sequelae of other nontraumatic intracranial haemorrhage
G683.00	Sequelae of cerebral infarction
G68W.00	Sequelae/other + unspecified cerebrovascular diseases
G68X.00	Sequelae of stroke,not specfd as h'morrhage or infarction
G6W..00	Cereb infarct due unsp occlus/stenos precerebr arteries
G6X..00	Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
G6y..00	Other specified cerebrovascular disease
G6z..00	Cerebrovascular disease NOS

Read codes used to identify cases of CKD

Dialysis codes

Read code	Description
14V2.11	H/O: kidney dialysis
7L1A.00	Compensation for renal failure
7L1A.11	Dialysis for renal failure
7L1A000	Renal dialysis
7L1A011	Thomas intravascular shunt for dialysis
7L1A100	Peritoneal dialysis
7L1A200	Haemodialysis NEC
7L1Ay00	Other specified compensation for renal failure
7L1Az00	Compensation for renal failure NOS
7L1B.11	Placement ambulatory dialysis apparatus - compens renal fail
7L1B000	Insertion of ambulatory peritoneal dialysis catheter
7L1B100	Removal of ambulatory peritoneal dialysis catheter
7L1By00	Placement ambulatory apparatus- compensate renal failure OS
7L1Bz00	Placement ambulatory apparatus- compensate renal failure NOS
7L1C.00	Placement other apparatus for compensation for renal failure
7L1C000	Insertion of temporary peritoneal dialysis catheter
7L1Cy00	Placement other apparatus- compensate for renal failure OS
7L1Cz00	Placement other apparatus- compensate for renal failure NOS
8882.00	Intestinal dialysis
SP01500	Mechanical complication of dialysis catheter
SP05613	[X] Peritoneal dialysis associated peritonitis

TA02.00	Accid cut,puncture,perf,h'ge - kidney dialysis/oth perfusion
TA02000	Accid cut,puncture,perf,h'ge - kidney dialysis
TA02011	Accidental cut/puncture/perf/haem'ge during renal dialysis
TA12000	Foreign object left in body during kidney dialysis
TA12011	Foreign object left in body during renal dialysis
TA22000	Failure of sterile precautions during kidney dialysis
TA22011	Failure of sterile precautions during renal dialysis
TA42000	Mechanical failure of apparatus during kidney dialysis
TA42011	Mechanical failure of apparatus during renal dialysis
TB11.00	Kidney dialysis with complication, without blame
TB11.11	Renal dialysis with complication, without blame
U641.00	[X]Kidny dialysis caus abn reac pt/lat comp no misad at time
Z1A..00	Dialysis training
Z1A1.00	Peritoneal dialysis training
Z1A1.11	PD - Peritoneal dialysis training
Z1A2.00	Haemodialysis training
Z1A2.11	HD - Haemodialysis training
Z919.00	Care of haemodialysis equipment
Z919100	Priming haemodialysis lines
Z919200	Washing back through haemodialysis lines
Z919300	Reversing haemodialysis lines
Z919400	Recirculation of the dialysis machine
Z91A.00	Peritoneal dialysis bag procedure
Z91A100	Putting additive into peritoneal dialysis bag
ZV45100	[V]Renal dialysis status
ZV56.00	[V]Aftercare involving intermittent dialysis
ZV56000	[V]Aftercare involving extracorporeal dialysis

ZV56011	[V]Aftercare involving renal dialysis NOS
ZV56100	[V]Preparatory care for dialysis
ZV56y00	[V]Other specified aftercare involving intermittent dialysis
ZV56y11	[V]Aftercare involving peritoneal dialysis
ZV56z00	[V]Unspecified aftercare involving intermittent dialysis
ZVu3G00	[X]Other dialysis

CKD codes

Read code	Description
K05..00	Chronic renal failure
K05..11	Chronic uraemia
K05..12	End stage renal failure
K050.00	End stage renal failure
K06..00	Renal failure unspecified
K06..11	Uraemia NOS
K060.00	Renal impairment
K060.11	Impaired renal function

Appendix 7

Validation of method used to estimate baseline clinical values

Table A7.1 Validation of method used to estimate baseline clinical values: comparison of simple mean and multilevel model results

	Cases	Observed and estimated baseline mean value (SD)			F-test			
		Observed	Mean value model	Multilevel model	RSS for mean value model	RSS for multilevel model	df for multilevel model	F-value (p)
SBP	3643	146 (21)	142 (14)	146 (14)	849746	648467	40	28 (<0.0001)
HbA_{1C}	3643	8.6 (2.4)	7.3 (1.2)	8.0 (1.6)	17187	7205	40	125 (<0.0001)
BMI	3643	30.8 (6.0)	30.3 (5.7)	30.5 (5.7)	8590	4170	39	98 (<0.0001)
Total cholesterol	3643	5.7 (1.3)	5.0 (0.9)	5.5 (1.0)	4835	1656	40	173 (<0.0001)
eGFR	3618	72 (17)	71 (16)	71 (15)	212206	126256	40	61 (<0.0001)

Note: Cohort restricted to cases with observed clinical value within 90 days of diagnosis of diabetes. RSS = residual sum of squares. df = degrees of freedom

Graphs used to compare observed and modelled risk factors

Figure A7.1 Observed and modelled BMI over study period

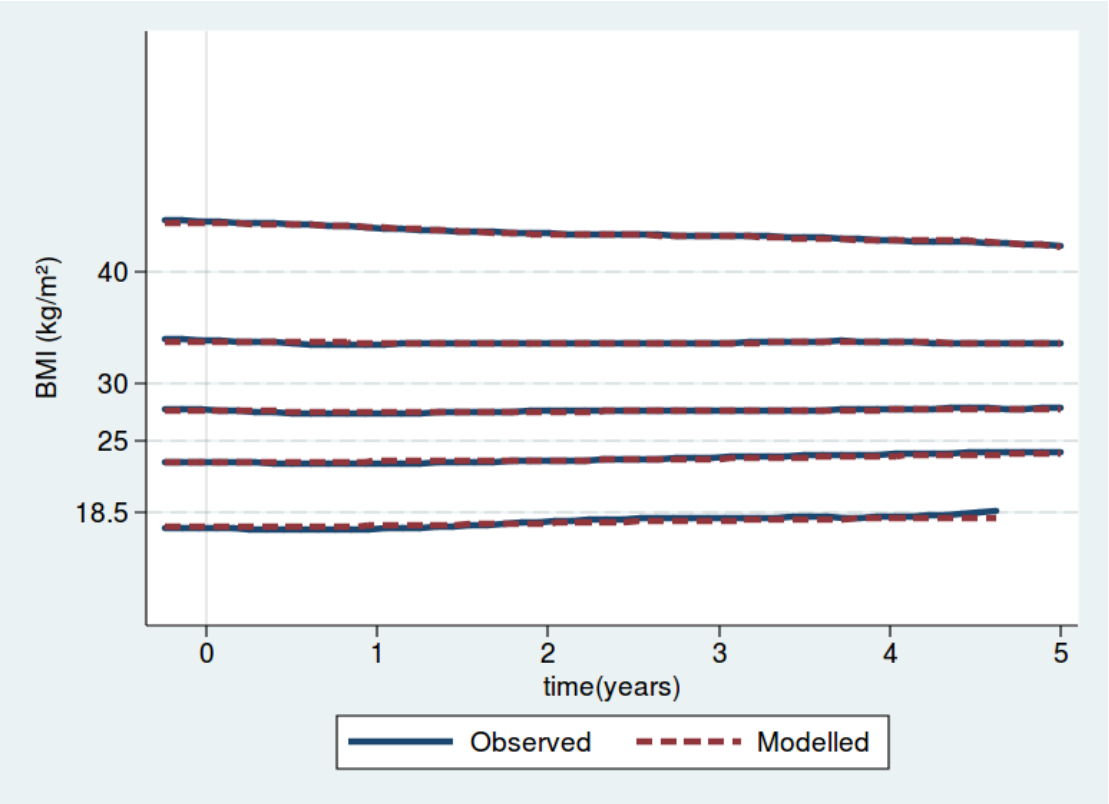


Figure A7.2 Observed and modelled total cholesterol over study period

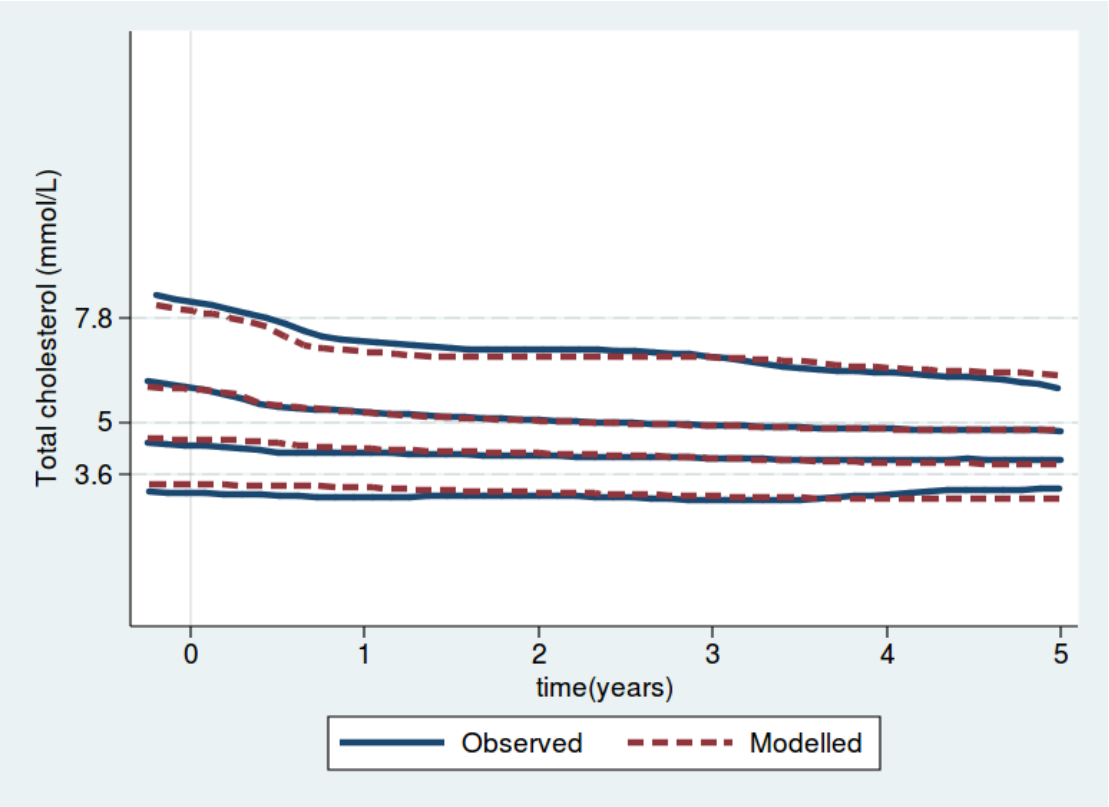
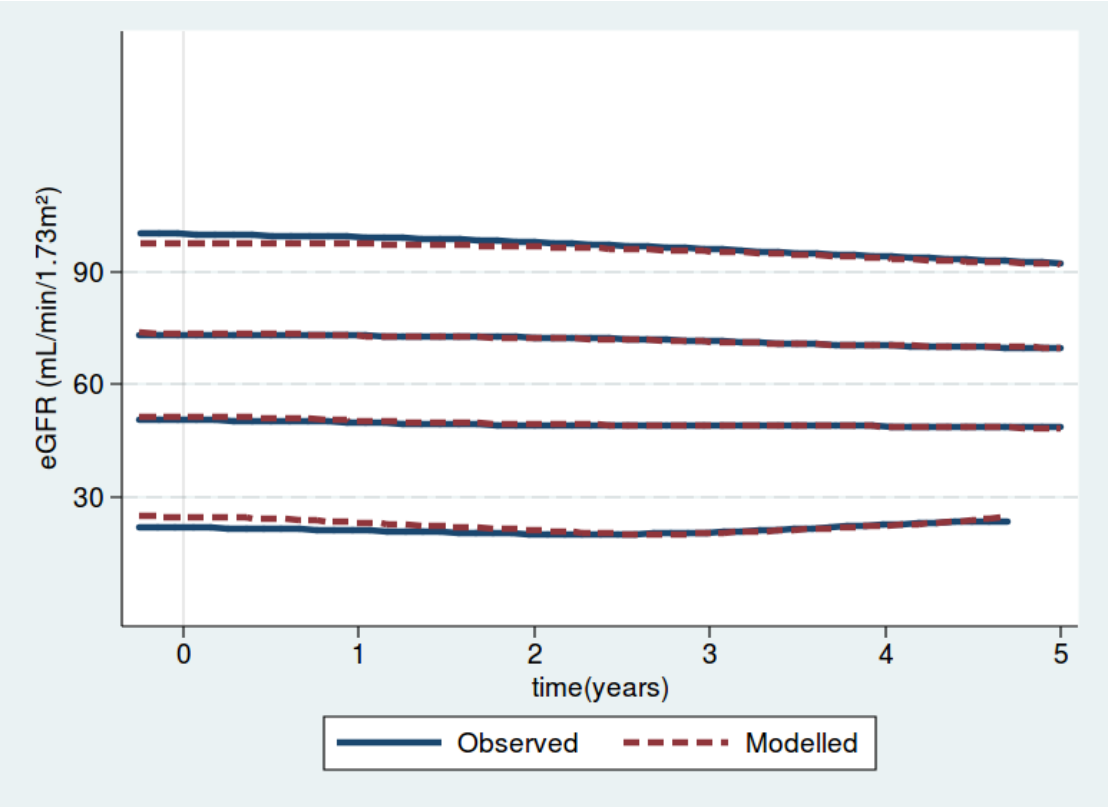
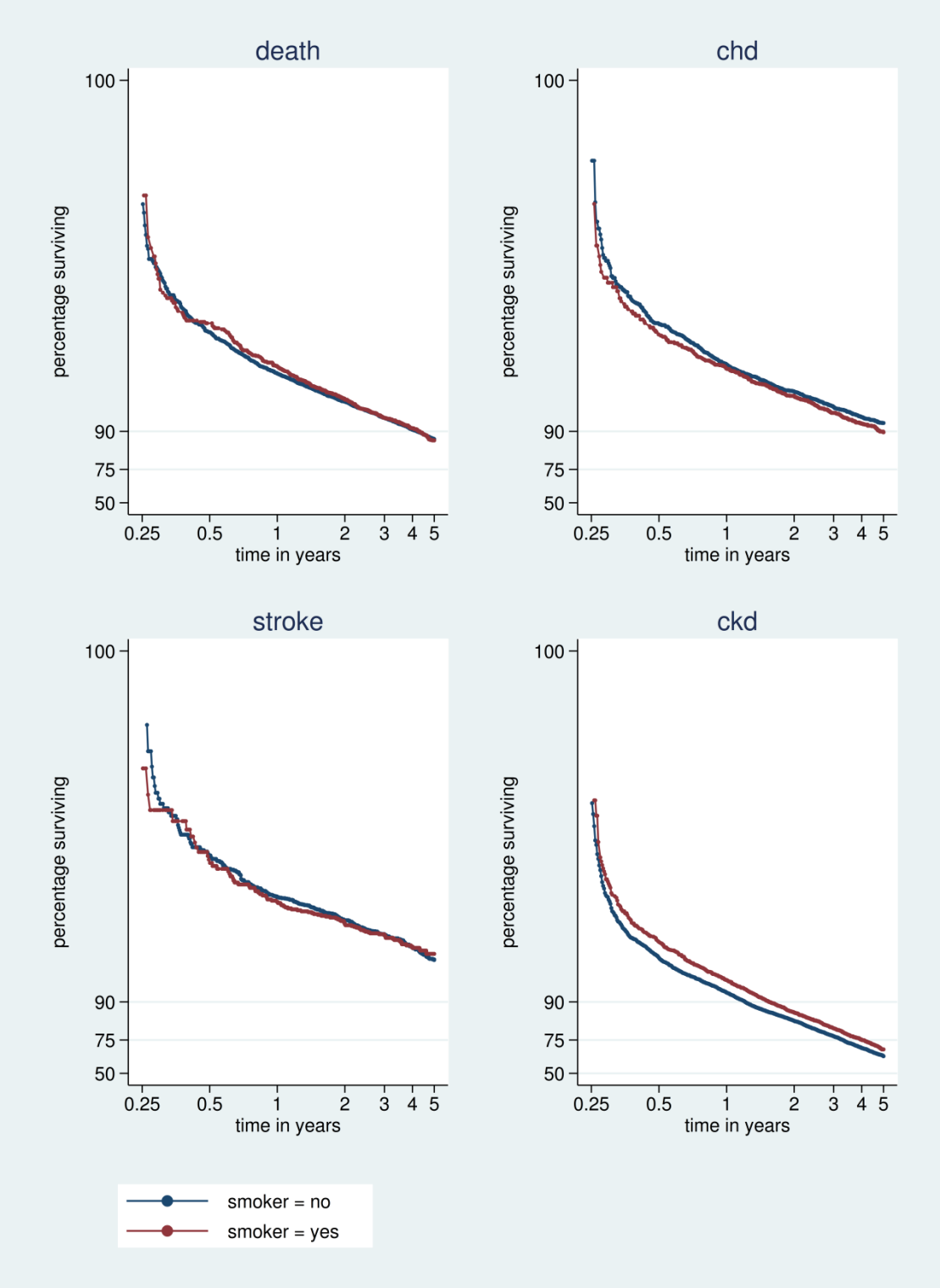


Figure A7.3 Observed and modelled eGFR over study period



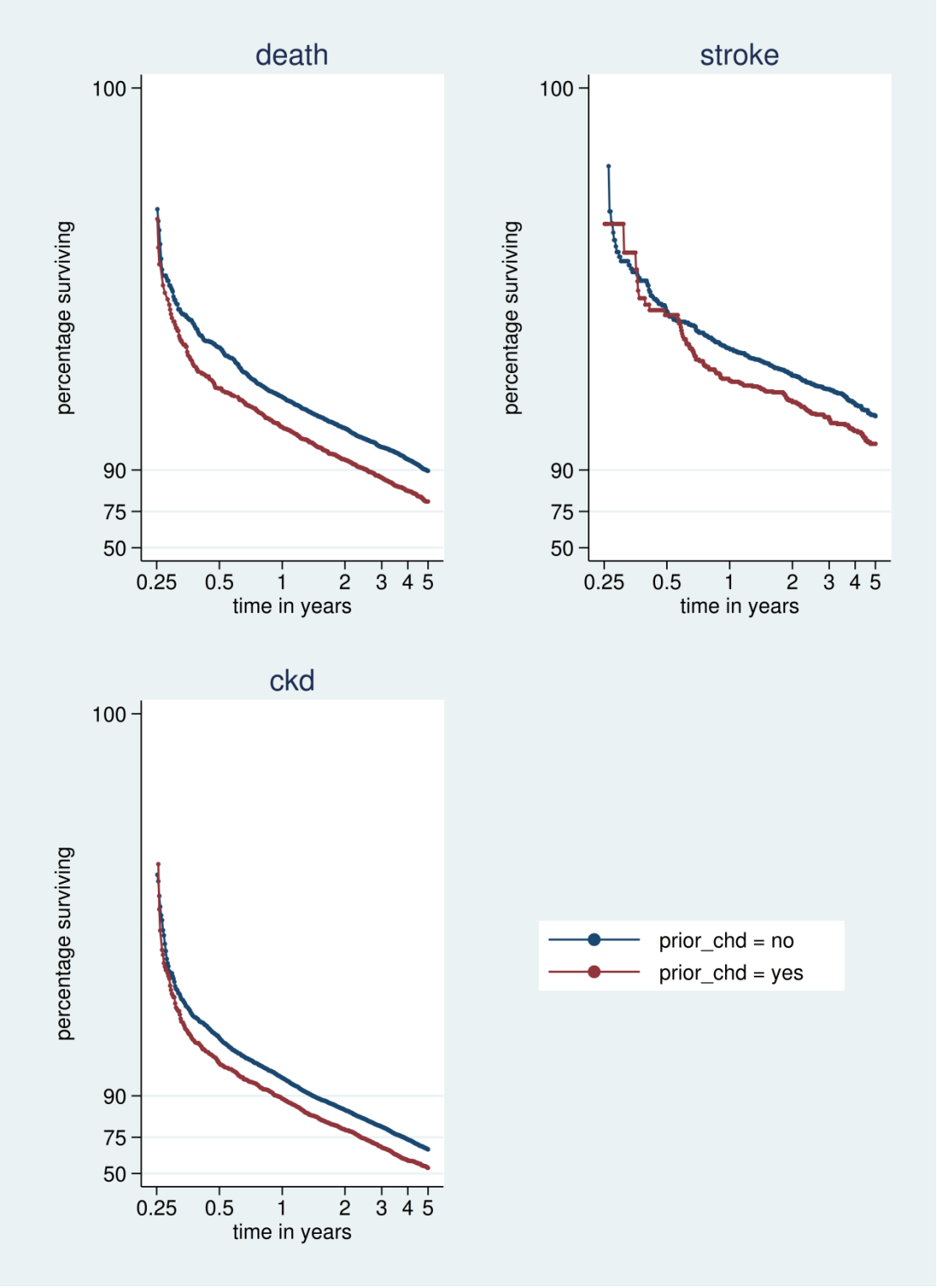
Log-log plots used to assess PH assumption for each prediction model

Figure A7.4 Log-log plots: smoker at diagnosis of diabetes



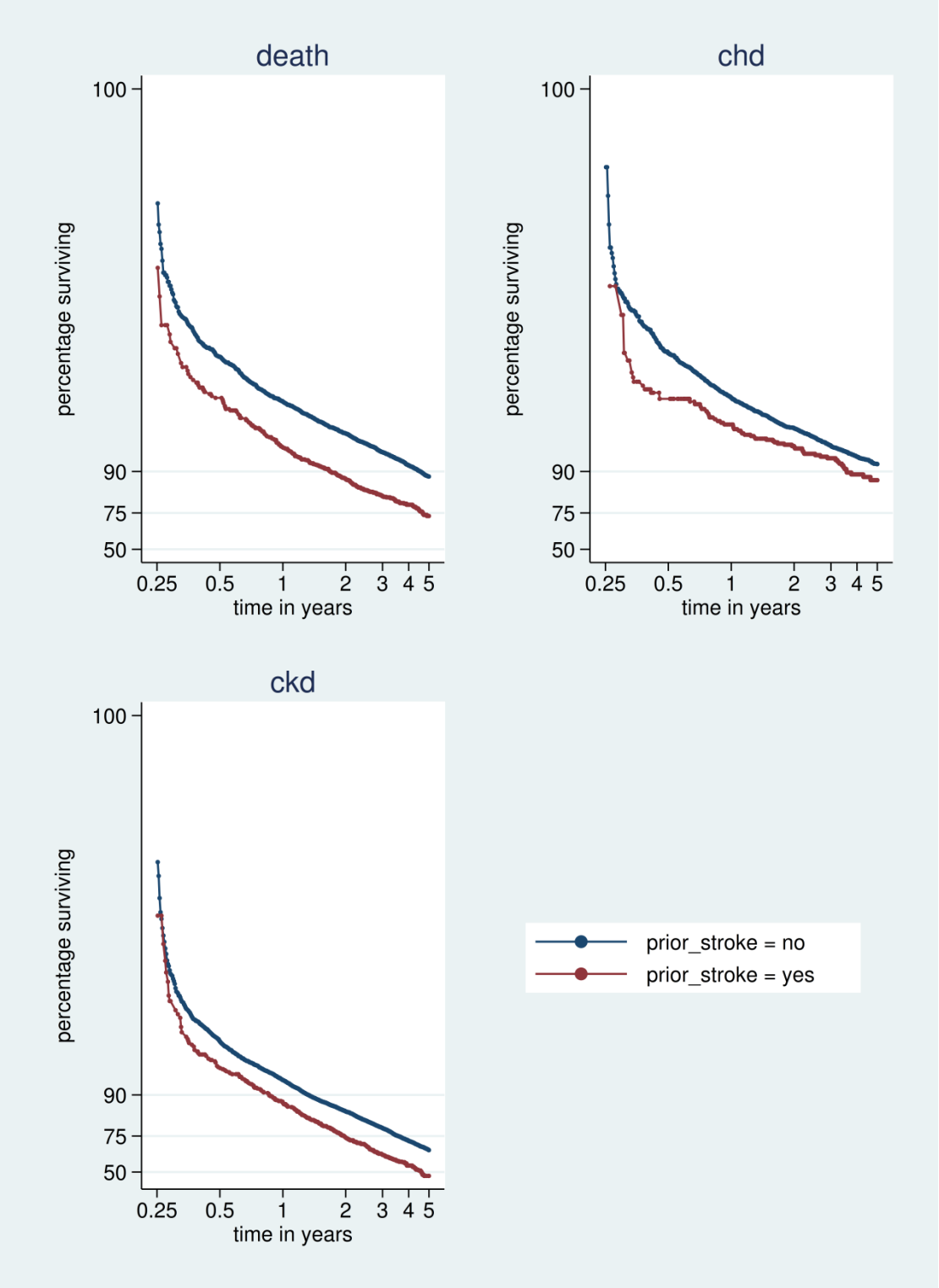
Note: Both axes are on logarithmic scales.

Figure A7.5 Log-log plots: CHD prior to diabetes or in first 3 months following diabetes



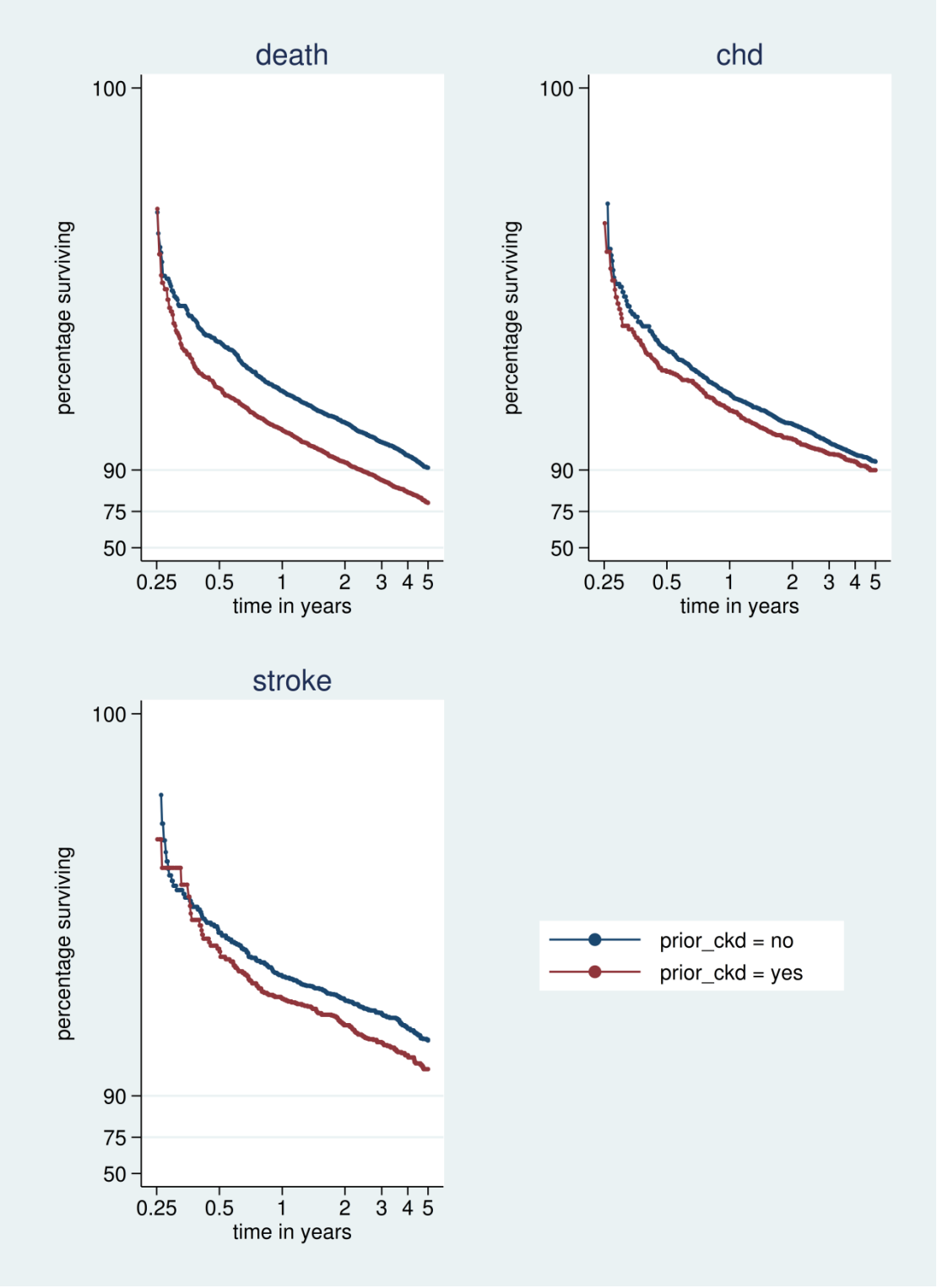
Note: Both axes are on logarithmic scales.

Figure A7.6 Log-log plots: stroke prior to diabetes or in first 3 months following diabetes



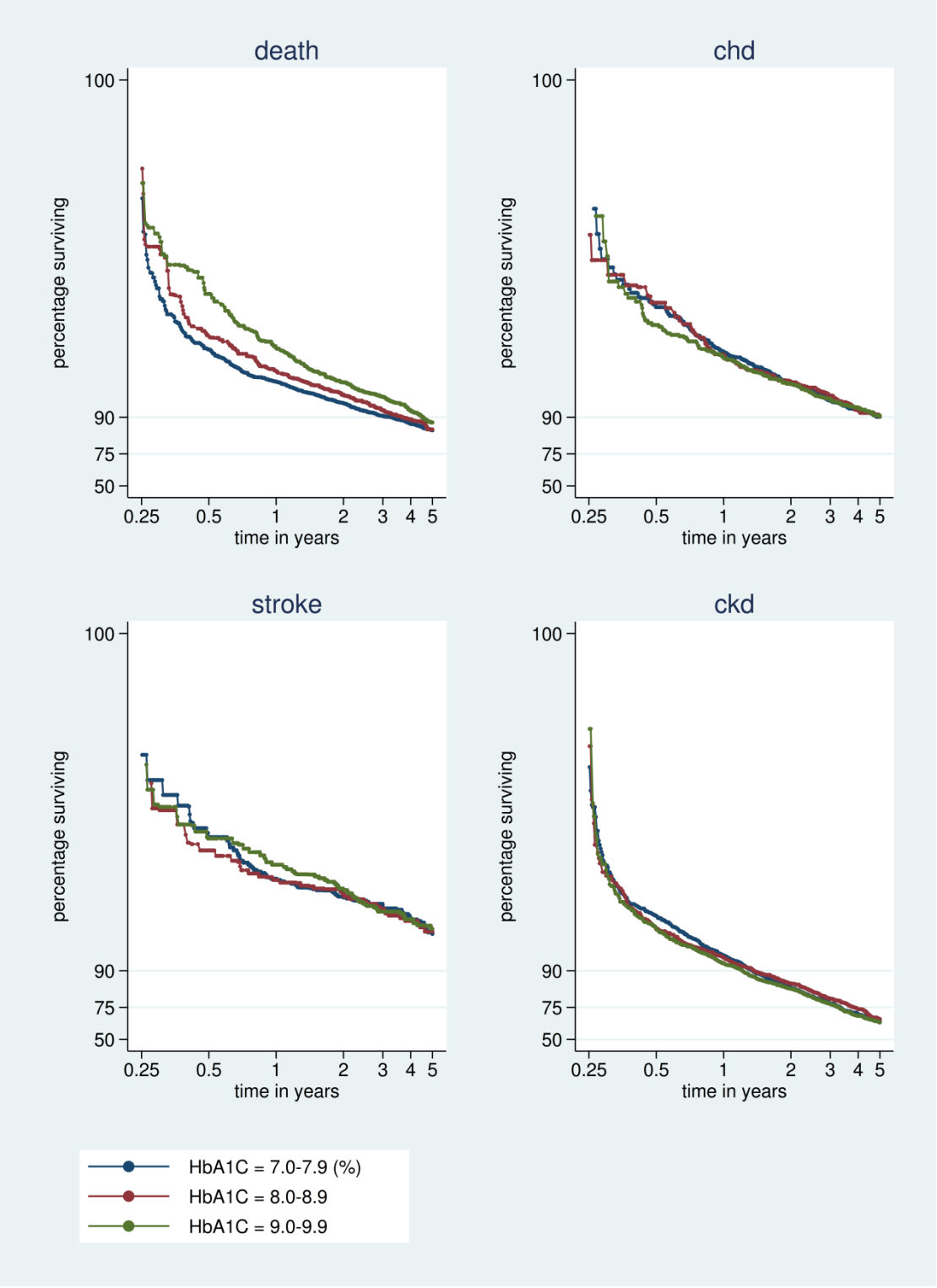
Note: Both axes are on logarithmic scales.

Figure A7.7 Log-log plots: CKD prior to diabetes or in first 3 months following diabetes



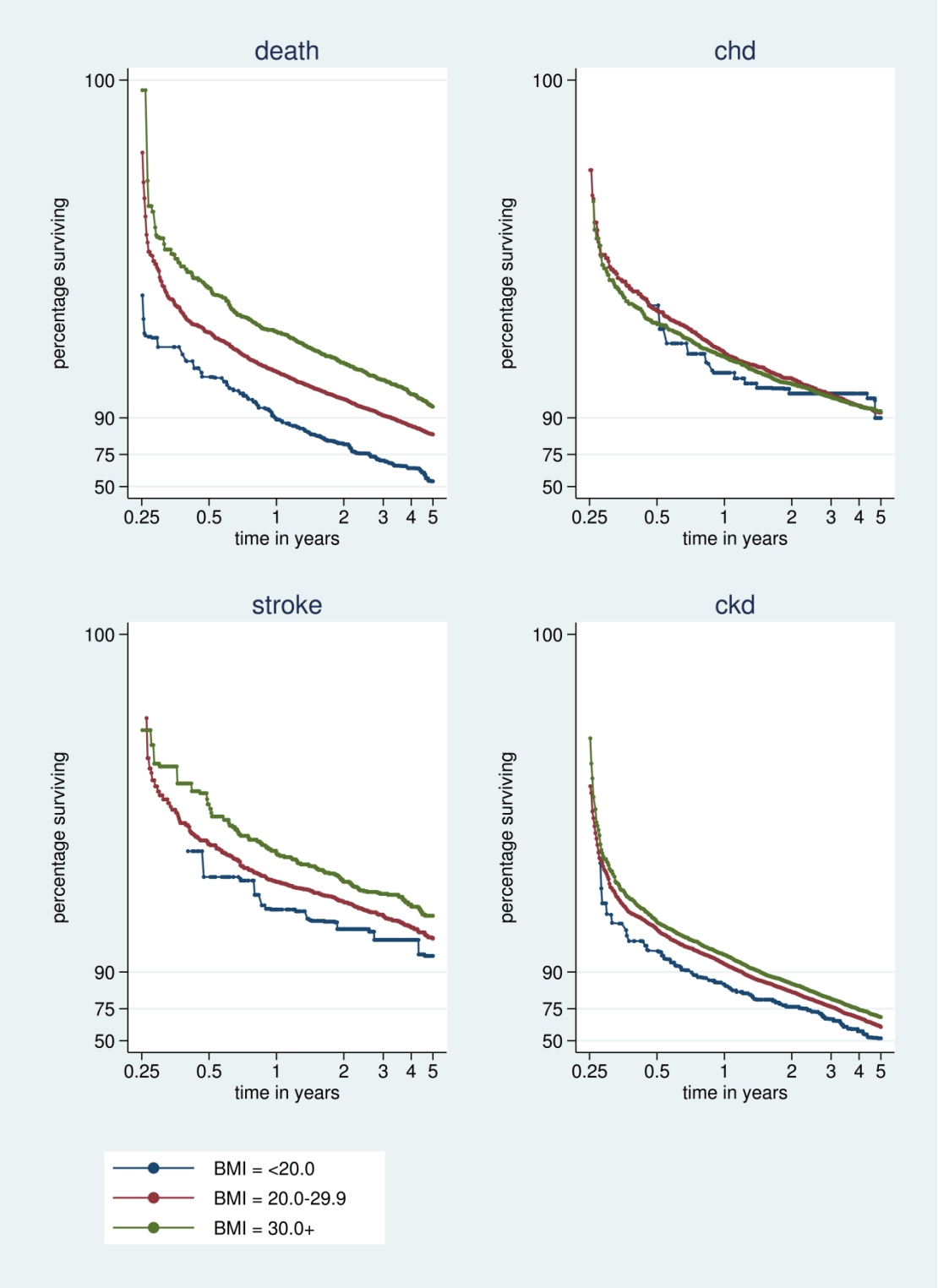
Note: Both axes are on logarithmic scales.

Figure A7.8 Log-log plots: HbA_{1C}



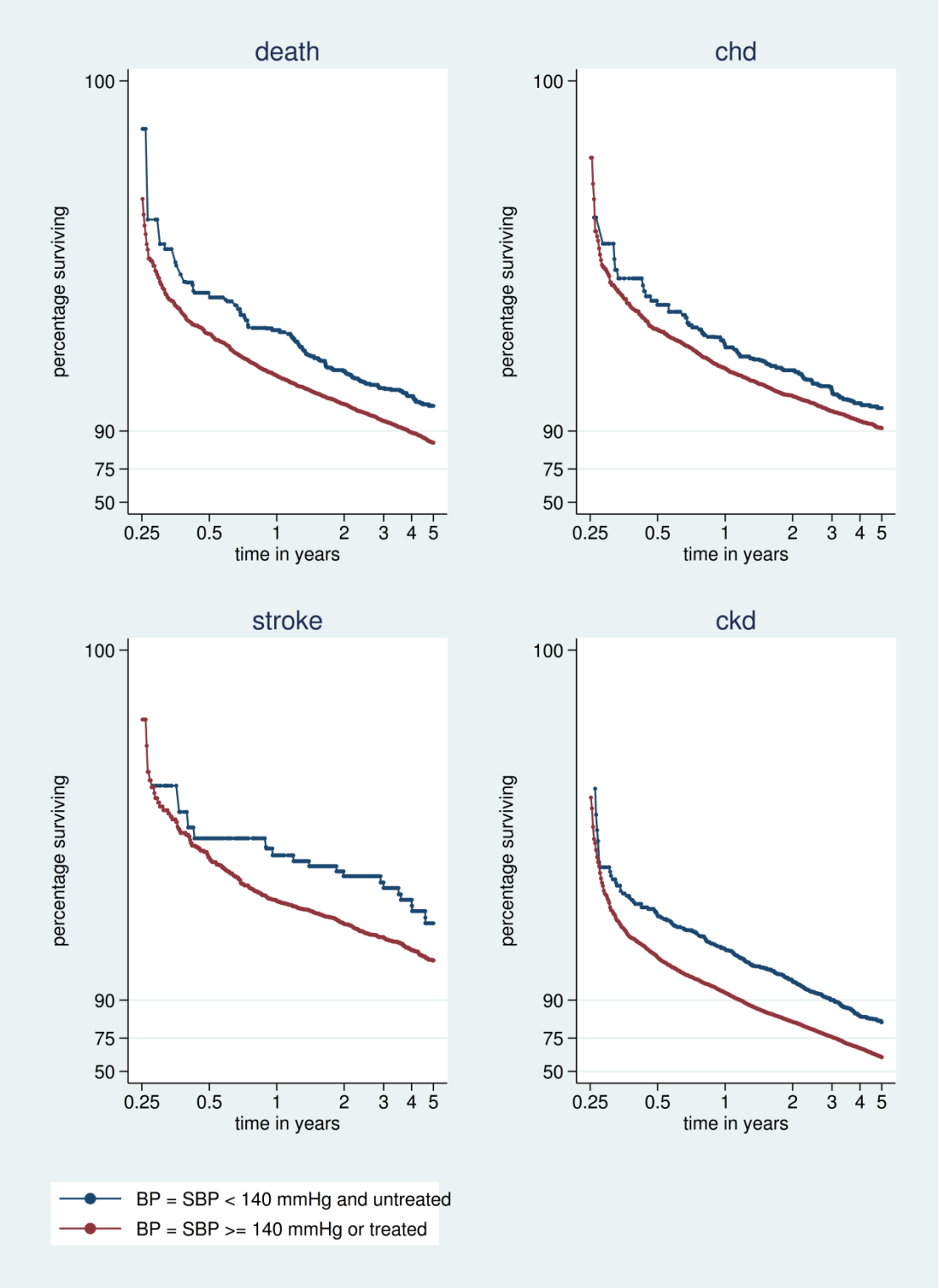
Note: Both axes are on logarithmic scales.

Figure A7.9 Log-log plots: BMI



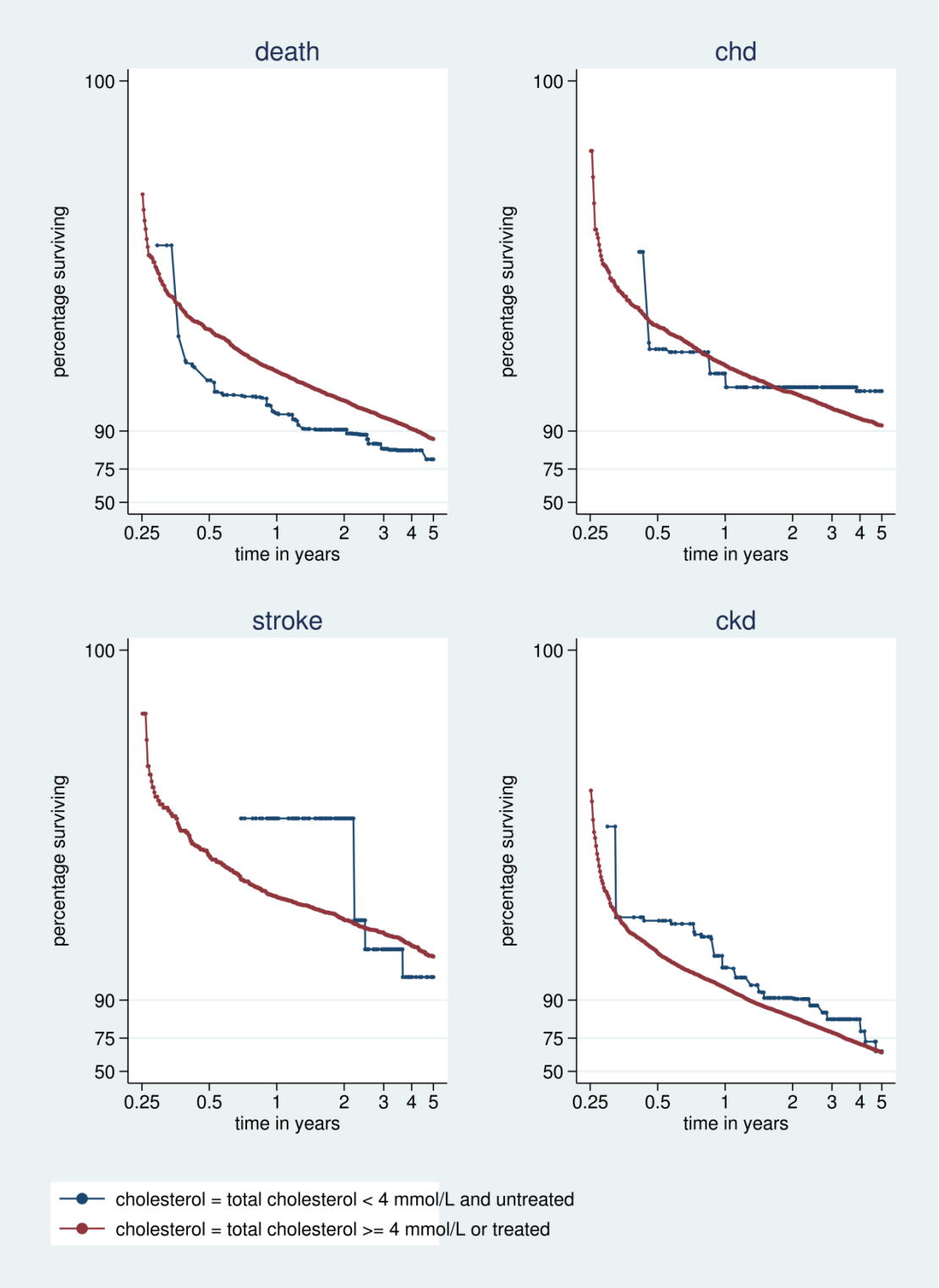
Note: Both axes are on logarithmic scales.

Figure A7.10 Log-log plots: High SBP or treated BP compared with low and untreated BP



Note: Both axes are on logarithmic scales.

Figure A7.11 Log-log plots: High or treated cholesterol compared with low and untreated cholesterol



Note: Both axes are on logarithmic scales.

Management of diabetes and diabetic-related risks

Table A7.2 Percentage of smokers continuing to smoke following diabetes diagnosis

Years following diagnosis	Percentage still smoking
0	100%
1	99%
2	94%
3	80%
4	63%
5	47%

Table A7.3 Percentage of cases prescribed or using drugs of interest before and after diabetes diagnosis

	One year prior to diagnosis	At diagnosis	3 months after diagnosis	One year after diagnosis
Diabetes management				
Any oral antidiabetic	-	5%	32%	46%
Biguanide	-	3%	21%	34%
Sulphonylurea	-	2%	14%	21%
Glitazone	-	0%	6%	23%
Meglitinide	-	0%	2%	6%
Acarbose / Guarnine	-	0%	1%	2%
Insulin	-	-	3%	3%
Prevention of cardiovascular disease				
Blood pressure lowering drugs	58%	64%	69%	74%
Lipid lowering drugs	14%	19%	29%	42%
Aspirin (prescribed)	23%	28%	34%	41%
Aspirin (OTC)	2%	4%	6%	8%

Results of multilevel models to predict baseline clinical values

Table A7.4 Multilevel model used to estimate baseline systolic blood pressure

Systolic BP		coefficient	p	95% CI	
year of diagnosis	1998	0.34	0.525	-0.72	1.41
(reference year = 2000)	1999	0.07	0.889	-0.86	0.99
(1998 and 2001 omitted due	2001	-1.20	0.001	-1.93	-0.46
to collinearity)	2002	-1.79	<0.001	-2.48	-1.10
	2003	-3.14	<0.001	-3.81	-2.46
age at diagnosis	35-44	-10.10	<0.001	-10.96	-9.23
(reference age group = 55-64)	45-54	-3.32	<0.001	-3.94	-2.70
	65-74	2.95	<0.001	2.42	3.47
	75-84	4.92	<0.001	4.28	5.57
	85-94	1.88	0.002	0.67	3.09
	95+	-10.37	0.001	-16.42	-4.32
male		-1.95	<0.001	-2.36	-1.54
smoker		-0.28	0.069	-0.59	0.02
Townsend quintile	(least deprived) 1	-0.06	0.817	-0.52	0.41
(reference quintile = 3)	2	-0.14	0.561	-0.60	0.33
	4	-0.42	0.072	-0.88	0.04
	(most deprived) 5	-0.55	0.030	-1.04	-0.05
region	north	-0.21	0.391	-0.70	0.27
(reference = middle)	south	-0.04	0.887	-0.52	0.45
comorbidities	prior chd	-1.40	<0.001	-1.91	-0.90
	prior chd	2.07	<0.001	1.62	2.51
	prior stroke	1.32	0.001	0.57	2.06
drug treatments	insulin	2.52	<0.001	1.56	3.49
	sulphonylurea	0.40	0.005	0.12	0.67
	biguanide	-0.14	0.216	-0.36	0.08
	acarbose	-1.40	0.167	-3.37	0.58
	meglitinide	1.22	0.061	-0.06	2.49
	glitazone	-0.06	0.824	-0.59	0.47
	statin	-1.50	<0.001	-1.71	-1.28
	other lipid lowering	-0.83	0.022	-1.53	-0.12
	antianginal(excl. CCB)	-1.84	<0.001	-2.26	-1.41
	aspirin	-0.01	0.954	-0.25	0.24
	OTC aspirin	-9.58	0.104	-21.13	1.97
	other antiplatelet	-0.88	0.006	-1.50	-0.25
	angiotensin-II receptor antagonist	-1.11	<0.001	-1.46	-0.76
	ACE inhibitor	-2.93	<0.001	-3.15	-2.71
	alphablocker	-1.67	<0.001	-2.05	-1.29
	calcium channel blocker	-2.42	<0.001	-2.71	-2.13
	diuretic	-2.44	<0.001	-2.72	-2.15
slope(per day)		-0.004	<0.001	-0.004	-0.003
intercept		150.89	<0.001	150.05	151.74

Table A7.5 Multilevel model used to estimate baseline BMI

BMI		coefficient	p	95% CI	
year of diagnosis	1998	-1.09	<0.001	-1.46	-0.72
(reference year = 2000)	1999	0.78	<0.001	0.47	1.09
(1998 and 2001 omitted due	2001	0.00	0.983	-0.25	0.25
to collinearity)	2002	-0.66	<0.001	-0.90	-0.42
	2003	-0.18	0.141	-0.41	0.06
age at diagnosis	35-44	3.04	<0.001	2.74	3.33
(reference age group = 55-64)	45-54	1.02	<0.001	0.82	1.23
	65-74	-1.88	<0.001	-2.05	-1.70
	75-84	-4.23	<0.001	-4.45	-4.01
	85-94	-6.25	<0.001	-6.73	-5.76
	95+	-8.36	<0.001	-11.52	-5.20
male		-1.28	<0.001	-1.42	-1.15
smoker		-0.23	<0.001	-0.27	-0.18
Townsend quintile	(least deprived) 1	-0.69	<0.001	-0.86	-0.53
(reference quintile = 3)	2	0.18	0.033	0.01	0.34
	4	0.28	0.001	0.12	0.44
	(most deprived) 5	0.66	<0.001	0.49	0.83
region	north	0.18	0.051	0.00	0.36
(reference = middle)	south	-0.22	0.014	-0.40	-0.05
comorbidities	prior chd	-0.07	0.382	-0.23	0.09
	prior chd	0.19	0.014	0.04	0.34
	prior stroke	-0.01	0.969	-0.28	0.27
drug treatments	insulin	1.06	<0.001	0.94	1.18
	sulphonylurea	0.57	<0.001	0.53	0.60
	biguanide	-0.28	<0.001	-0.30	-0.25
	acarbose	-0.56	<0.001	-0.82	-0.31
	meglitinide	0.14	0.105	-0.03	0.30
	glitazone	0.67	<0.001	0.61	0.74
	statin	-0.08	<0.001	-0.11	-0.05
	other lipid lowering	-0.09	0.073	-0.19	0.01
	antianginal(excl. CCB)	0.12	0.001	0.05	0.18
	aspirin	0.03	0.150	-0.01	0.06
	OTC aspirin	1.28	0.041	0.05	2.50
	other antiplatelet	-0.15	0.002	-0.25	-0.05
	angiotensin-II receptor antagonist	0.21	<0.001	0.15	0.27
	ACE inhibitor	-0.19	<0.001	-0.22	-0.15
	alphablocker	0.38	<0.001	0.32	0.45
	calcium channel blocker	0.06	0.031	0.01	0.11
	diuretic	0.08	0.003	0.03	0.13
slope(per day)		-0.0003	<0.0010	-0.0003	-0.0002
intercept		32.04	<0.0010	31.75	32.33

Table A7.6 Multilevel model used to estimate baseline total cholesterol

Total cholesterol		coefficient	p	95% CI	
year of diagnosis	1998	0.065	0.061	-0.003	0.133
(reference year = 2000)	1999	-0.058	0.050	-0.116	0.000
(1998 and 2001 omitted due	2001	-0.118	0.000	-0.163	-0.072
to collinearity)	2002	-0.199	0.000	-0.242	-0.157
	2003	-0.263	0.000	-0.305	-0.222
age at diagnosis	35-44	-0.061	0.022	-0.114	-0.009
(reference age group = 55-64)	45-54	0.034	0.076	-0.004	0.072
	65-74	-0.092	0.000	-0.124	-0.059
	75-84	-0.213	0.000	-0.253	-0.173
	85-94	-0.398	0.000	-0.481	-0.316
	95+	-0.630	0.018	-1.152	-0.108
male		-0.369	0.000	-0.395	-0.344
smoker		0.071	0.000	0.051	0.091
Townsend quintile	(least deprived) 1	0.012	0.401	-0.016	0.041
(reference quintile = 3)	2	0.039	0.008	0.010	0.067
	4	-0.001	0.970	-0.029	0.028
	(most deprived) 5	0.007	0.674	-0.024	0.037
region	north	0.054	0.000	0.024	0.084
(reference = middle)	south	0.003	0.830	-0.027	0.033
comorbidities	prior chd	0.051	0.002	0.019	0.084
	prior chd	0.045	0.002	0.017	0.072
	prior stroke	0.022	0.349	-0.024	0.069
drug treatments	insulin	-0.074	0.019	-0.135	-0.012
	sulphonylurea	-0.008	0.377	-0.027	0.010
	biguanide	-0.046	0.000	-0.061	-0.030
	acarbose	0.143	0.027	0.017	0.269
	meglitinide	0.016	0.699	-0.065	0.097
	glitazone	0.117	0.000	0.084	0.150
	statin	-1.162	0.000	-1.176	-1.148
	other lipid lowering	-0.126	0.000	-0.168	-0.084
	antianginal(excl. CCB)	0.036	0.016	0.007	0.065
	aspirin	-0.012	0.152	-0.029	0.005
	OTC aspirin	0.100	0.320	-0.097	0.297
	other antiplatelet	-0.128	0.000	-0.172	-0.084
	angiotensin-II receptor antagonist	-0.010	0.478	-0.036	0.017
	ACE inhibitor	-0.041	0.000	-0.058	-0.024
	alphablocker	-0.106	0.000	-0.135	-0.076
	calcium channel blocker	0.062	0.000	0.040	0.083
	diuretic	0.052	0.000	0.031	0.073
slope(per day)		-0.0003	0.000	-0.0004	-0.0003
time(CDF)		0.505	0.000	0.476	0.534
intercept		5.999	0.000	5.946	6.051

Table A7.7 Multilevel model used to estimate baseline eGFR

eGFR		coefficient	p	95% CI	
year of diagnosis	1998	1.55	0.000	0.74	2.37
(reference year = 2000)	1999	0.79	0.026	0.09	1.49
(1998 and 2001 omitted due	2001	-0.34	0.215	-0.89	0.20
to collinearity)	2002	-0.43	0.099	-0.94	0.08
	2003	-0.57	0.025	-1.07	-0.07
age at diagnosis	35-44	5.07	0.000	4.43	5.71
(reference age group = 55-64)	45-54	2.24	0.000	1.78	2.69
	65-74	-2.92	0.000	-3.30	-2.53
	75-84	-5.73	0.000	-6.20	-5.26
	85-94	-10.12	0.000	-10.98	-9.26
	95+	-14.07	0.000	-18.17	-9.96
male		2.39	0.000	2.09	2.70
smoker		0.68	0.000	0.46	0.91
Townsend quintile	(least deprived) 1	-0.24	0.170	-0.58	0.10
(reference quintile = 3)	2	-0.49	0.005	-0.83	-0.15
	4	0.11	0.506	-0.22	0.45
	(most deprived) 5	-0.17	0.345	-0.53	0.19
region	north	-0.46	0.011	-0.82	-0.11
(reference = middle)	south	-0.28	0.124	-0.63	0.08
comorbidities	prior chd	-1.45	0.000	-1.83	-1.08
	prior chd	-19.36	0.000	-19.69	-19.03
	prior stroke	-1.20	0.000	-1.75	-0.66
drug treatments	insulin	-1.85	0.000	-2.57	-1.13
	sulphonylurea	-0.76	0.000	-0.96	-0.55
	biguanide	0.63	0.000	0.46	0.79
	acarbose	-1.07	0.179	-2.63	0.49
	meglitinide	-1.43	0.004	-2.40	-0.45
	glitazone	-0.65	0.001	-1.02	-0.28
	statin	0.07	0.376	-0.09	0.23
	other lipid lowering	-3.76	0.000	-4.27	-3.24
	antianginal(excl. CCB)	-0.39	0.018	-0.71	-0.07
	aspirin	0.64	0.000	0.45	0.83
	OTC aspirin	1.12	0.305	-1.02	3.25
	other antiplatelet	-0.55	0.024	-1.03	-0.07
	angiotensin-II receptor antagonist	-0.81	0.000	-1.10	-0.52
	ACE inhibitor	-0.49	0.000	-0.66	-0.31
	alphablocker	-0.95	0.000	-1.28	-0.62
	calcium channel blocker	0.03	0.825	-0.21	0.26
	diuretic	-1.56	0.000	-1.79	-1.33
slope(per day)		-0.001	0.000	-0.001	-0.001
intercept		81.62	0.000	81.00	82.25

REFERENCES

1. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. World Health Organization; 2006.

2. International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. Diabetes Care. 2009;32(7):1327-34.

3. World Health Organization. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. Diabetes Research and Clinical Practice. 2011;93(3):299-309.

4. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2010;33 Suppl 1:S62-9.

5. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for diabetes development. Lancet. 2012;379(9833):2279-90.

6. Ryan R, Newnham A, Khunti K, Majeed A. New cases of diabetes mellitus in England and Wales, 1994-1998: database study. Public Health. 2005;119(10):892-9.

7. Newnham A, Ryan R.; Khunti, K.; Majeed, A. Prevalence of diagnosed diabetes in general practice in England and Wales, 1994 to 1998. Health Statistics Quarterly. 2002;14:5-13.

8. Imkampe AK, Gulliford MC. Increasing socio-economic inequality in type 2 diabetes prevalence--repeated cross-sectional surveys in England 1994-2006. European Journal of Public Health. 2011;21(4):484-90.

9. Goyder EC, McNally PG, Drucquer M, Spiers N, Botha JL. Shifting of care for diabetes from secondary to primary care, 1990-5: review of general practices. BMJ. 1998;316(7143):1505-6.

10. Khunti K, Ganguli S. Who looks after people with diabetes: primary or secondary care? Journal of the Royal Society of Medicine. 2000;93(4):183-6.

11. Stone MA, Wilkinson JC, Charpentier G, Clochard N, Grassi G, Lindblad U, et al. Evaluation and comparison of guidelines for the management of people with type 2 diabetes from eight European countries. Diabetes Research and Clinical Practice. 2010;87(2):252-60.

12. Kontopantelis E, Reeves D, Valderas JM, Campbell S, Doran T. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. BMJ Quality and Safety. 2013;22(1):53-64.

13. Alshamsan R, Millett C, Majeed A, Khunti K. Has pay for performance improved the management of diabetes in the United Kingdom? Primary Care Diabetes. 2010;4(2):73-8.

14. McGovern MP, Williams DJ, Hannaford PC, Taylor MW, Lefevre KE, Boroujerdi MA, et al. Introduction of a new incentive and target-based contract for family physicians in the UK: good for older patients with diabetes but less good for women? Diabetic Medicine. 2008;25(9):1083-9.

15. Millett C, Gray J, Saxena S, Netuveli G, Majeed A. Impact of a pay-for-performance incentive on support for smoking cessation and on smoking prevalence among people with diabetes. Canadian Medical Association Journal. 2007;176(12):1705-10.

16. Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. Quality of primary care in England with the introduction of pay for performance. The New England Journal of Medicine. 2007;357(2):181-90.

17. Berry JD, Dyer A, Cai X, Garside DB, Ning H, Thomas A, et al. Lifetime risks of cardiovascular disease. The New England Journal of Medicine. 2012;366(4):321-9.

18. Thomas RL, Dunstan F, Luzio SD, Roy Chowdury S, Hale SL, North RV, et al. Incidence of diabetic retinopathy in people with type 2 diabetes mellitus attending the Diabetic Retinopathy Screening Service for Wales: retrospective analysis. BMJ. 2012;344:e874.

19. Hill CJ, Fogarty DG. Changing trends in end-stage renal disease due to diabetes in the United kingdom. Journal of Renal Care. 2012;38 Suppl 1:12-22.

20. Kannel WB, McGee DL. Diabetes and cardiovascular disease. The Framingham study. The Journal of the American Medical Association. 1979;241(19):2035-8.

21. Girman CJ, Kou TD, Brodovicz K, Alexander CM, O'Neill EA, Engel S, et al. Risk of acute renal failure in patients with Type 2 diabetes mellitus. Diabetic Medicine. 2012;29(5):614-21.

22. Vamos EP, Bottle A, Edmonds ME, Valabhji J, Majeed A, Millett C. Changes in the incidence of lower extremity amputations in individuals with and without diabetes in England between 2004 and 2008. Diabetes Care. 2010;33(12):2592-7.

23. Hemmingsen B, Lund SS, Gluud C, Vaag A, Almdal T, Hemmingsen C, et al. Intensive glycaemic control for patients with type 2 diabetes: systematic review with meta-analysis and trial sequential analysis of randomised clinical trials. BMJ. 2011;343:d6898.

24. Butalia S, Leung AA, Ghali WA, Rabi DM. Aspirin effect on the incidence of major adverse cardiovascular events in patients with diabetes mellitus: a systematic review and meta-analysis. Cardiovascular diabetology. 2011;10:25.

25. Zhang CY, Sun AJ, Zhang SN, Wu CN, Fu MQ, Xia G, et al. Effects of intensive glucose control on incidence of cardiovascular events in patients with type 2 diabetes: a meta-analysis. Annals of medicine. 2010;42(4):305-15.

26. Eeg-Olofsson K, Cederholm J, Nilsson PM, Zethelius B, Nunez L, Gudbjornsdottir S, et al. Risk of cardiovascular disease and mortality in overweight and obese patients with type 2 diabetes: an observational study in 13,087 patients. Diabetologia. 2009;52(1):65-73.

27. Holman RR, Paul SK, Bethel MA, Neil HA, Matthews DR. Long-term follow-up after tight control of blood pressure in type 2 diabetes. The New England Journal of Medicine. 2008;359(15):1565-76.

28. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HA. 10-year follow-up of intensive glucose control in type 2 diabetes. The New England Journal of Medicine. 2008;359(15):1577-89.

29. Gaede P, Lund-Andersen H, Parving HH, Pedersen O. Effect of a multifactorial intervention on mortality in type 2 diabetes. The New England Journal of Medicine. 2008;358(6):580-91.

30. Bolen S, Feldman L, Vassy J, Wilson L, Yeh HC, Marinopoulos S, et al. Systematic review: comparative effectiveness and safety of oral medications for type 2 diabetes mellitus. Annals of internal medicine. 2007;147(6):386-99.

31. Selvin E, Coresh J, Golden SH, Brancati FL, Folsom AR, Steffes MW. Glycemic control and coronary heart disease risk in persons with and without diabetes: the atherosclerosis risk in communities study. Archives of internal medicine. 2005;165(16):1910-6.

32. Dailey G. Early and intensive therapy for management of hyperglycemia and cardiovascular risk factors in patients with type 2 diabetes. Clinical therapeutics. 2011;33(6):665-78.

33. Davies MJ, Heller S, Skinner TC, Campbell MJ, Carey ME, Cradock S, et al. Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomised controlled trial. BMJ. 2008;336(7642):491-5.

34. Gaede P, Vedel P, Larsen N, Jensen GV, Parving HH, Pedersen O. Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. The New England Journal of Medicine. 2003;348(5):383-93.

35. Spencer EA, Pirie KL, Stevens RJ, Beral V, Brown A, Liu B, et al. Diabetes and modifiable risk factors for cardiovascular disease: the prospective Million Women Study. European journal of epidemiology. 2008;23(12):793-9.

36. De Backer G, Ambrosioni E, Borch-Johnsen K, Brotons C, Cifkova R, Dallongeville J, et al. European guidelines on cardiovascular disease prevention in clinical practice. Third Joint Task Force of European and Other Societies on Cardiovascular Disease Prevention in Clinical Practice. European heart journal. 2003;24(17):1601-10.

37. British Cardiac Society, British Hypertension Society, Diabetes UK, Heart UK, Primary Care Cardiovascular Society, Stroke Association. JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. Heart. 2005;91 Suppl 5:v1-52.

38. National Institute for Health and Clinical Excellence (NICE). Clinical Guidelines and Evidence Review for Lipid Modification: cardiovascular risk assessment and the primary and secondary prevention of cardiovascular disease (CG67). Available from: http://www.nice.org.uk/cg67. [Accessed: 30/7/2013]

39. National Institute for Health and Clinical Excellence (NICE). Type 2 diabetes (partially updated by CG87) (CG66). Available from: http://guidance.nice.org.uk/CG66. [Accessed: 03/10/2013]

40. Law MR, Wald NJ, Morris JK, Jordan RE. Value of low dose combination treatment with blood pressure lowering drugs: analysis of 354 randomised trials. BMJ. 2003;326(7404):1427.

41. Tonelli M, Lloyd A, Clement F, Conly J, Husereau D, Hemmelgarn B, et al. Efficacy of statins for primary prevention in people at low cardiovascular risk: a meta-analysis. Canadian Medical Association Journal. 2011;183(16):E1189-202.

42. Cholesterol Treatment Trialists C, Mihaylova B, Emberson J, Blackwell L, Keech A, Simes J, et al. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. Lancet. 2012;380(9841):581-90.

43. National Institute for Health and Clinical Excellence (NICE). Type 2 Diabetes - newer agents (partial update of CG66) (CG87). Available from: http://guidance.nice.org.uk/CG87. [Accessed: 8/9/2012]

44. Yudkin JS, Chaturvedi N. Developing risk stratification charts for diabetic and nondiabetic subjects. Diabetic Medicine. 1999;16(3):219-27.

45. Stevens RJ, Kothari V, Adler AI, Stratton IM, United Kingdom Prospective Diabetes Study G. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). Clinical science. 2001;101(6):671-9.

46. Kothari V, Stevens RJ, Adler AI, Stratton IM, Manley SE, Neil HA, et al. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. Stroke; a journal of cerebral circulation. 2002;33(7):1776-81.

47. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Systematic reviews. 2012;1:10.

48. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart (British Cardiac Society). 2012;98(5):360-9.

49. Chamnan P, Simmons RK, Sharp SJ, Griffin SJ, Wareham NJ. Cardiovascular risk assessment scores for people with diabetes: a systematic review. Diabetologia. 2009;52(10):2001-14.

50. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of clinical epidemiology. 2013;66(3):268-77.

51. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS medicine. 2012;9(11):e1001344.

52. Mukamal KJ, Kizer JR, Djousse L, Ix JH, Zieman S, Siscovick DS, et al. Prediction and classification of cardiovascular disease risk in older adults with diabetes. Diabetologia. 2013;56(2):275-83.

53. Kengne AP, Patel A, Marre M, Travert F, Lievre M, Zoungas S, et al. Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. European journal of cardiovascular prevention and rehabilitation. 2011;18(3):393-8.

54. Davis WA, Knuiman MW, Davis TM. An Australian cardiovascular risk equation for type 2 diabetes: the Fremantle Diabetes Study. Internal medicine journal. 2010;40(4):286-92.

55. Elley CR, Robinson E, Kenealy T, Bramley D, Drury PL. Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the new zealand diabetes cohort study. Diabetes Care. 2010;33(6):1347-52.

56. Cederholm J, Eeg-Olofsson K, Eliasson B, Zethelius B, Nilsson PM, Gudbjornsdottir S, et al. Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. Diabetes Care. 2008;31(10):2038-43.

57. Yang X, So WY, Kong AP, Ma RC, Ko GT, Ho CS, et al. Development and validation of a total coronary heart disease risk score in type 2 diabetes mellitus. The American journal of cardiology. 2008;101(5):596-601.

58. Yang X, So WY, Kong AP, Ho CS, Lam CW, Stevens RJ, et al. Development and validation of stroke risk equation for Hong Kong Chinese patients with type 2 diabetes: the Hong Kong Diabetes Registry. Diabetes Care. 2007;30(1):65-70.

59. Donnan PT, Donnelly L, New JP, Morris AD. Derivation and validation of a prediction score for major coronary heart disease events in a U.K. type 2 diabetic population. Diabetes Care. 2006;29(6):1231-6.

60. Folsom AR, Chambless LE, Duncan BB, Gilbert AC, Pankow JS, Atherosclerosis Risk in Communities Study I. Prediction of coronary heart disease in middle-aged adults with diabetes. Diabetes Care. 2003;26(10):2777-84.

61. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. American heart journal. 1991;121(1 Pt 2):293-8.

62. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart. 2007;93(2):172-6.

63. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ. 2007;341:c6624.

64. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335(7611):136.

65. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation. 2002;105(3):310-5.

66. Jardine MJ, Hata J, Woodward M, Perkovic V, Ninomiya T, Arima H, et al. Prediction of kidney-related outcomes in patients with type 2 diabetes. American journal of kidney diseases. 2012;60(5):770-8.

67. O'Seaghdha CM, Lyass A, Massaro JM, Meigs JB, Coresh J, D'Agostino RB, Sr., et al. A risk score for chronic kidney disease in the general population. The American journal of medicine. 2012;125(3):270-7.

68. Alssema M, Newson RS, Bakker SJ, Stehouwer CD, Heymans MW, Nijpels G, et al. One risk assessment tool for cardiovascular disease, type 2 diabetes, and chronic kidney disease. Diabetes Care. 2012;35(4):741-8.

69. Hanratty R, Chonchol M, Havranek EP, Powers JD, Dickinson LM, Ho PM, et al. Relationship between blood pressure and incident chronic kidney disease in hypertensive patients. Clinical journal of the American Society of Nephrology. 2011;6(11):2605-11.

70. Halbesma N, Jansen DF, Heymans MW, Stolk RP, de Jong PE, Gansevoort RT, et al. Development and validation of a general population renal risk score. Clinical journal of the American Society of Nephrology. 2011;6(7):1731-8.

71. Hippisley-Cox J, Coupland C. Predicting the risk of chronic Kidney Disease in men and women in England and Wales: prospective derivation and external validation of the QKidney Scores. BMC family practice. 2010;11:49.

72. Hanratty R, Chonchol M, Miriam Dickinson L, Beaty BL, Estacio RO, Mackenzie TD, et al. Incident chronic kidney disease and the rate of kidney function decline in individuals with hypertension. Nephrology, dialysis, transplantation. 2010;25(3):801-7.

73. Chien KL, Lin HJ, Lee BC, Hsu HC, Lee YT, Chen MF. A prediction model for the risk of incident chronic kidney disease. The American journal of medicine. 2010;123(9):836-46 e2.

74. Kshirsagar AV, Bang H, Bomback AS, Vupputuri S, Shoham DA, Kern LM, et al. A simple algorithm to predict incident kidney disease. Archives of internal medicine. 2008;168(22):2466-73.

75. Fox CS, Larson MG, Leip EP, Culleton B, Wilson PW, Levy D. Predictors of new-onset kidney disease in a community-based population. The Journal of the American Medical Association. 2004;291(7):844-50.

76. Xu L, Chan WM, Hui YF, Lam TH. Association between HbA1c and cardiovascular disease mortality in older Hong Kong Chinese with diabetes. Diabetic Medicine. 2012;29(3):393-8.

77. Skriver MV, Stovring H, Kristensen JK, Charles M, Sandbaek A. Short-term impact of HbA1c on morbidity and all-cause mortality in people with type 2 diabetes: a Danish population-based observational study. Diabetologia. 2012;55(9):2361-70.

78. Andersson C, van Gaal L, Caterson ID, Weeke P, James WP, Coutinho W, et al. Relationship between HbA1c levels and risk of cardiovascular adverse outcomes and all-cause mortality in overweight and obese cardiovascular high-risk women and men with type 2 diabetes. Diabetologia. 2012;55(9):2348-55.

79. Kerr D, Partridge H, Knott J, Thomas PW. HbA1c 3 months after diagnosis predicts premature mortality in patients with new onset type 2 diabetes. Diabetic Medicine. 2011;28(12):1520-4.

80. Currie CJ, Peters JR, Tynan A, Evans M, Heine RJ, Bracco OL, et al. Survival as a function of HbA(1c) in people with type 2 diabetes: a retrospective cohort study. Lancet. 2010;375(9713):481-9.

81. Clarke PM, Gray AM, Briggs A, Farmer AJ, Fenn P, Stevens RJ, et al. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). Diabetologia. 2004;47(10):1747-59.

82. Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. BMC Medical Research Methodology. 2009;9:34.

83. Department of Health. National service framework for diabetes: standards 2001. Available from: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4002951. [Accessed: 8/9/2012]

84. Department of Health. QOF Business Ruleset for Diabetes: version 22. Available from: http://www.pcc.nhs.uk/uploads/QOF/business%20rules%20v22.0/diabetes_ruleset_v22_0.pdf. [Accessed: 8/9/2012]

85. GPcontract website. List of UK general practices. Available from: http://gpcontract.co.uk/#childorgs. [Accessed: 04/10/2013]

86. Office for National Statistics. Key Health Statistics from General Practice 1998. London; 2000.

87. Elliott P, Peakman TC, Biobank UK. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. International Journal of Epidemiology. 2008;37(2):234-44.

88. Study UPD. UK Prospective Diabetes Study (UKPDS). VIII. Study design, progress and performance. Diabetologia. 1991;34(12):877-90.

89. Department of Health. Recommendations for the provision of services in primary care for people with diabetes: 2005 Available from : http://www.diabetes.org.uk/Documents/Professionals/primary_recs.pdf. [Accessed: 8/9/2012]

90. Ryan R, Majeed A. Prevalence of ischaemic heart disease and its management with statins and aspirin in general practice in England and Wales, 1994-1998. Health Statistics Quarterly. 2001;12:34-9.

91. HealthKnowledge. Sources of routine mortality and morbidity data, including primary care data, and how they are collected and published at international, national, regional and district levels.. Available from: http://www.healthknowledge.org.uk/public-health-textbook/health-information/3b-sickness-health/collection-routine-ad-hoc-data. [Accessed: 4/10/2013]

92. Khunti K, Baker R, Rumsey M, Lakhani M. Quality of care of patients with diabetes: collation of data from multi-practice audits of diabetes in primary care. Family Practice. 1999;16(1):54-9.

93. Wells S, Benett I, Holloway G, Harlow V. Area-wide diabetes Care: the Manchester experience with primary health care teams 1991-1997. Diabetic Medicine. 1998;15 Suppl 3:S49-53.

94. Whitford DL, Southern AJ, Braid E, Roberts SH. Comprehensive diabetes care in North Tyneside. Diabetic Medicine. 1995;12(8):691-5.

95. Whitford DL, Roberts SH. Changes in prevalence and site of care of diabetes in a health district 1991-2001. Diabetic Medicine. 2004;21(6):640-3.

96. Morris AD, Boyle DI, MacAlpine R, Emslie-Smith A, Jung RT, Newton RW, et al. The diabetes audit and research in Tayside Scotland (DARTS) study: electronic record linkage to create a diabetes register. DARTS/MEMO Collaboration. BMJ. 1997;315(7107):524-8.

97. Burnett SD, Woolf CM, Yudkin JS. Developing a district diabetic register. BMJ. 1992;305(6854):627-30.

98. Cunningham S, McAlpine R, Leese G, Brennan G, Sullivan F, Connacher A, et al. Using web technology to support population-based diabetes care. Journal of Diabetes Science and Technology. 2011;5(3):523-34.

99. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

100. NICE. NICE Guidelines: Chronic kidney disease (CG73)(updated 8/9/2012). Available from: http://www.nice.org.uk/CG73. [Accessed: 04/10/2013]

101. Marshall T, Lancashire R, Sharp D, Peters TJ, Cheng KK, Hamilton W. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. Gut. 2011;60(9):1242-8.

102. Damery S, Nichols L, Holder R, Ryan R, Wilson S, Warmington S, et al. Assessing the predictive value of HIV indicator conditions in general practice: a case-control study using the THIN database. The British Journal of General Practice. 2013;63(611):e370-7.

103. Cegedim Medical Research UK. THIN Data Guide for Researchers 2011 version 2.2. Cegedim Medical Research UK; 2011.

104. Marshall T, Caley M, Hemming K, Gill P, Gale N, Jolly K. Mixed methods evaluation of targeted case finding for cardiovascular disease prevention using a stepped wedged cluster RCT. BMC Public Health. 2012;12:908.

105. Marshall T, Westerby P, Chen J, Fairfield M, Harding J, Westerby R, et al. The Sandwell Project: a controlled evaluation of a programme of targeted screening for prevention of cardiovascular disease in primary care. BMC Public Health. 2008;8:73.

106. The Clinical Practice Research Datalink (formerly known as the General Practice Research database). Available from: http://www.cprd.com/. [Accessed 8/8/2012]

107. The Health Improvement Network database (THIN). Available from: http://csdmruk.cegedim.com/. [Accessed 8/8/2012]

108. QResearch. The QResearch database. Available from: http://www.qresearch.org. [Accessed 8/8/2012]

109. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.

110. Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine. 2000;19(4):453-73.

111. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. The British Journal of General Practice. 2010;60(572):e128-36.

112. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346:f2350.

113. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010;29(28):2920-31.

114. Department of Health. The Good Practice Guidelines for GP electronic patient records - version 42012. Available from: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_125310. [Accesed: 8/8/2012]

115. Haynes K, Bilker WB, Tenhave TR, Strom BL, Lewis JD. Temporal and within practice variability in the health improvement network. Pharmacoepidemiology and Drug Safety. 2011;20(9):948-55.

116. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PloS One. 2012;7(1):e30412.

117. Carey IM, Cook DG, De WS, Bremner SA, Richards N, Caine S, et al. Developing a large electronic primary care database (Doctors' Independent Network) for research. International Journal of Medical Informatics. 2004;73(5):443-53.

118. Jordan K, Clarke AM, Symmons DP, Fleming D, Porcheret M, Kadam UT, et al. Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases. British Journal of General Practice. 2007;57(534):7-14.

119. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Medical Research Methodology. 2013;13(1):105.

120. Tate AR, Martin AG, Ali A, Cassell JA. Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational study using primary care records of patients with ovarian cancer. BMJ Open. 2011;1(1):e000025.

121. Nicholson A, Tate AR, Koeling R, Cassell JA. What does validation of cases in electronic record databases mean? The potential contribution of free text. Pharmacoepidemiology and Drug Safety. 2011;20(3):321-4.

122. Anandarajah S, Tai T, de LS, Stevens P, O'Donoghue D, Walker M, et al. The validity of searching routinely collected general practice computer data to identify patients with chronic kidney disease (CKD): a manual review of 500 medical records. Nephrology Dialysis Transplantation. 2005;20(10):2089-96.

123. Liljeqvist GT, Staff M, Puech M, Blom H, Torvaldsen S. Automated data extraction from general practice records in an Australian setting: trends in influenza-like illness in sentinel general practices and emergency departments. BMC Public Health. 2011;11:435.

124. Ruigomez A, Martin-Merino E, Rodriguez LA. Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). Pharmacoepidemiology and Drug Safety. 2010;19(6):579-85.

125. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. Family Practice. 2004;21(4):396-412.

126. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. BritishJournal of Clinical Pharmacology. 2010;69(1):4-14.

127. Holt TA, Stables D, Hippisley-Cox J, O'Hanlon S, Majeed A. Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records. The British Journal of General Practice. 2008;58(548):192-6.

128. de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabetic Medicine. 2010;27(2):203-9.

129. Sadek AR, van Vlymen J, Khunti K, de Lusignan S. Automated identification of miscoded and misclassified cases of diabetes from computer records. Diabetic Medicine. 2012;29(3):410-4.

130. Shah A, Martinez C. A comparison of the cause of death recorded in GPRD with national mortality statistics in England and Wales. Phamacoepidemiology and Drug Safety. 2004;13(Supplement):S2-S3.

131. Gulliford MC, Charlton J. Is relative mortality of type 2 diabetes mellitus decreasing? American Journal of Epidemiology. 2009;169(4):455-61.

132. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. Pharmacoepidemiology and Drug Safety. 2009;18(1):76-83.

133. Stratton IM, Adler AI, Neil HA, Matthews DR, Manley SE, Cull CA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. BMJ. 2000;321(7258):405-12.

134. Szatkowski L, Lewis S, McNeill A, Huang Y, Coleman T. Can data from primary care medical records be used to monitor national smoking prevalence? Journal of Epidemiology and Community Health. 2012;66(9):791-5.

135. Renal Association. eGFR Calculator. Available from: http://www.renal.org/eGFRcalc/GFR.pl. [Accessed: 6/8/2012]

136. Jones M KJ. Capturing ethnicity data in primary care: challenges and feasibility in a diverse metropolitan population. Diversity in Health and Social Care. 2007;4(3):221-10.

137. Hox JJ. Multilevel analysis : techniques and applications. Mahwah, N.J.: Lawrence Erlbaum Associates; 2002.

138. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge ; New York: Cambridge University Press; 2007.

139. Collett D. Modelling survival data in medical research. 1st ed. London ; New York: Chapman & Hall; 1994.

140. Cox NJ. Pweibull: probability plot for data versus fitted Weibull distribution. RePEc (Research Papers in Economics). Available from: http://fmwww.bc.edu/RePEc/bocode/p/pweibull.html. [Accessed: 8/9/2012]

141. Adams RJ, Chimowitz MI, Alpert JS, Awad IA, Cerqueria MD, Fayad P, et al. Coronary risk evaluation in patients with transient ischemic attack and ischemic stroke: a scientific statement for healthcare professionals from the Stroke Council and the Council on Clinical Cardiology of the American Heart Association/American Stroke Association. Circulation. 2003;108(10):1278-90.

142. Touze E, Varenne O, Chatellier G, Peyrard S, Rothwell PM, Mas JL. Risk of myocardial infarction and vascular death after transient ischemic attack and ischemic stroke: a systematic review and meta-analysis. Stroke. 2005;36(12):2748-55.

143. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Medical Ressearch Methodology. 2010;10:112.

144. Mulnier HE, Seaman HE, Raleigh VS, Soedamah-Muthu SS, Colhoun HM, Lawrenson RA. Mortality in people with type 2 diabetes in the UK. Diabetic Medicine. 2006;23(5):516-21.

145. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Statistics in Medicine. 1991;10(4):585-98.

146. Royston P, Carlin JB, White IR. Multiple imputation of missing values: New features for mim. Stata Journal. 2009;9(2):252-64.

147. Carlin JB, Sterne AC, White IR, Royston P, Kenward MG, Wood AM, et al. Multiple imputation needs to be used with care and reported in detail. Available from: http://www.bmj.com/rapid-response/2011/11/01/multiple-imputation-needs-be-used-care-and-reported-detail. [Accessed: 8/6/2012]

148. Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. Emerging Themes in Epidemiology. 2012;9(1):3.

149. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research. 1998;33(4):545-71.

150. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Medical Research Methodology. 2009;9:57.

151. UCLA Stata Library: Multiple Imputation Using ICE. Available from: http://www.ats.ucla.edu/STAT/stata/library/ice.htm. [Accessed: 10/7/2014]

152. White IR, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009;28(15):1982-98.

153. Nag S, Bilous R, Kelly W, Jones S, Roper N, Connolly V. All-cause and cardiovascular mortality in diabetic subjects increases significantly with reduced estimated glomerular filtration rate (eGFR): 10 years' data from the South Tees Diabetes Mortality study. Diabetic Medicine. 2007;24(1):10-7.

154. Gatling W, Guzder RN, Turnbull JC, Budd S, Mullee MA. The Poole Diabetes Study: how many cases of Type 2 diabetes are diagnosed each year during normal health care in a defined community? Diabetes Research and Clinical Practice. 2001;53(2):107-12.

155. Evans JM, Barnett KN, Ogston SA, Morris AD. Increasing prevalence of type 2 diabetes in a Scottish population: effect of increasing incidence or decreasing mortality? Diabetologia. 2007;50(4):729-32.

156. Guzder RN, Gatling W, Mullee MA, Byrne CD. Early mortality from the time of diagnosis of Type 2 diabetes: a 5-year prospective cohort study with a local age- and sex-matched comparison cohort. Diabetic Medicine. 2007;24(10):1164-7.

157. Royston P. Explained variation for survival models. Stata Journal. 2006;6(1):83-96.

158. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78(3):691-2.

159. Levesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ. 2010;340:b5087.

160. Feigin VL, Lawes CM, Bennett DA, Anderson CS. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. Lancet Neurology. 2003;2(1):43-53.

161. Harris MI, Klein R, Welborn TA, Knuiman MW. Onset of NIDDM occurs at least 4-7 yr before clinical diagnosis. Diabetes Care. 1992;15(7):815-9.

162. Echouffo-Tcheugui JB, Sargeant LA, Prevost AT, Williams KM, Barling RS, Butler R, et al. How much might cardiovascular disease risk be reduced by intensive therapy in people with screen-detected diabetes? Diabetic Medicine. 2008;25(12):1433-9.

163. National Institute for Health and Clinical Excellence (NICE). Preventing type 2 diabetes: risk identification and interventions for individuals at high risk (PH38). Available from: http://www.nice.org.uk/guidance/PH38. [Accessed: 10/7/2014]

164. Simmons RK, Echouffo-Tcheugui JB, Sharp SJ, Sargeant LA, Williams KM, Prevost AT, et al. Screening for type 2 diabetes and population mortality over 10 years (ADDITION-Cambridge): a cluster-randomised controlled trial. Lancet. 2012;380(9855):1741-8.

165. NHS Health Check website. Available from: http://www.healthcheck.nhs.uk/. [Accessed: 10/07/2014]

166. Carnethon MR, De Chavez PJ, Biggs ML, Lewis CE, Pankow JS, Bertoni AG, et al. Association of weight status with mortality in adults with incident diabetes. The Journal of the American Medical Association. 2012;308(6):581-90.

167. Perry JR, Voight BF, Yengo L, Amin N, Dupuis J, Ganser M, et al. Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. PLoS Genetics. 2012;8(5):e1002741.

168. Bombelli M, Facchetti R, Sega R, Carugo S, Fodri D, Brambilla G, et al. Impact of body mass index and waist circumference on the long-term risk of diabetes mellitus, hypertension, and cardiac organ damage. Hypertension. 2011;58(6):1029-35.

169. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabetic Medicine. 2010;27(8):887-95.

170. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes/metabolism Research and Reviews. 2000;16(3):164-71.

171. Hu FB, Stampfer MJ, Haffner SM, Solomon CG, Willett WC, Manson JE. Elevated risk of cardiovascular disease prior to clinical diagnosis of type 2 diabetes. Diabetes Care. 2002;25(7):1129-34.

172. Taubert G, Winkelmann BR, Schleiffer T, Marz W, Winkler R, Gok R, et al. Prevalence, predictors, and consequences of unrecognized diabetes mellitus in 3266 patients scheduled for coronary angiography. American Heart Journal. 2003;145(2):285-91.

173. Vancheri F, Curcio M, Burgio A, Salvaggio S, Gruttadauria G, Lunetta MC, et al. Impaired glucose metabolism in patients with acute stroke and no previous diagnosis of diabetes mellitus. QJM: An International Journal of Medicine. 2005;98(12):871-8.

174. Rathmann W, Icks A, Haastert B, Giani G, Lowel H, Mielck A. Undiagnosed diabetes mellitus among patients with prior myocardial infarction. Zeitschrift fur Kardiologie. 2002;91(8):620-5.

175. Turnbull F, Blood Pressure Lowering Treatment Trialists C. Effects of different blood-pressure-lowering regimens on major cardiovascular events: results of prospectively-designed overviews of randomised trials. Lancet. 2003;362(9395):1527-35.

176. Jesky M, Lambert A, Burden AC, Cockwell P. The impact of chronic kidney disease and cardiovascular comorbidity on mortality in a multiethnic population: a retrospective cohort study. BMJ Open. 2013;3(12):e003458.

177. McEwen LN, Kim C, Karter AJ, Haan MN, Ghosh D, Lantz PM, et al. Risk factors for mortality among patients with diabetes: the Translating Research Into Action for Diabetes (TRIAD) Study. Diabetes Care. 2007;30(7):1736-41.

178. Tolstrup JS, Hvidtfeldt UA, Flachs EM, Spiegelman D, Heitmann BL, Balter K, et al. Smoking and Risk of Coronary Heart Disease in Younger, Middle-Aged, and Older Adults. American Journal of Public Health. 2013;104(1):96-102.

179. Global Burden of Metabolic Risk Factors for Chronic Diseases C, Lu Y, Hajifathalian K, Ezzati M, Woodward M, Rimm EB, et al. Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: a pooled analysis of 97 prospective cohorts with 1.8 million participants. Lancet. 2014;383(9921):970-83.

180. Wannamethee SG, Shaper AG, Whincup PH, Lennon L, Sattar N. Impact of diabetes on cardiovascular disease risk and all-cause mortality in older men: influence of age at onset, diabetes duration, and established and novel risk factors. Archives of Internal Medicine. 2011;171(5):404-10.

181. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. Lancet. 1999;353(9147):89-92.

182. Turnbull F, Neal B, Algert C, Chalmers J, Chapman N, Cutler J, et al. Effects of different blood pressure-lowering regimens on major cardiovascular events in individuals with and without diabetes mellitus: results of prospectively designed overviews of randomized trials. Archives of Internal Medicine. 2005;165(12):1410-9.

183. Coresh J, Selvin E, Stevens LA, Manzi J, Kusek JW, Eggers P, et al. Prevalence of chronic kidney disease in the United States. The Journal of the American Medical Association. 2007;298(17):2038-47.

184. de Lusignan S, Tomson C, Harris K, van Vlymen J, Gallagher H. Creatinine fluctuation has a greater effect than the formula to estimate glomerular filtration rate on the prevalence of chronic kidney disease. Nephron Clinical Practice. 2011;117(3):c213-24.

185. Adler AI, Stratton IM, Neil HA, Yudkin JS, Matthews DR, Cull CA, et al. Association of systolic blood pressure with macrovascular and microvascular complications of type 2 diabetes (UKPDS 36): prospective observational study. BMJ. 2000;321(7258):412-9.

186. Hansson L, Zanchetti A, Carruthers SG, Dahlof B, Elmfeldt D, Julius S, et al. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. HOT Study Group. Lancet. 1998;351(9118):1755-62.

187. De Berardis G, Sacco M, Strippoli GF, Pellegrini F, Graziano G, Tognoni G, et al. Aspirin for primary prevention of cardiovascular events in people with diabetes: meta-analysis of randomised controlled trials. BMJ. 2009;339:b4531.

188. Heart Protection Study Collaborative G, Bulbulia R, Bowman L, Wallendszus K, Parish S, Armitage J, et al. Effects on 11-year mortality and morbidity of lowering LDL cholesterol with simvastatin for about 5 years in 20,536 high-risk individuals: a randomised controlled trial. Lancet. 2011;378(9808):2013-20.

189. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiology and Drug Safety. 2010;19(6):618-26.

190. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. Informatics in Primary Care. 2004;12(3):171-7.

191. Hall GC. Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiology and Drug Safety. 2009;18(2):120-31.

192. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in The Health Improvement Network. Pharmacoepidemiology and Drug Safety. 2009;18(8):730-6.

193. Kumarapeli P, Stepaniuk R, de Lusignan S, Williams R, Rowlands G. Ethnicity recording in general practice computer systems. Journal of Public Health. 2006;28(3):283-7.

194. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, Vanstaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. Journal of Public Health. 2013.

195. New Zealand Ministry of Health. Assessment and Management of Cardiovascular Risk. Available from: http://www.health.govt.nz/publication/assessment-and-management-cardiovascular-risk. [Accessed: 10/07/2014]

196. Price HC, Tucker L, Griffin SJ, Holman RR. The impact of individualised cardiovascular disease (CVD) risk estimates and lifestyle advice on physical activity in individuals at high risk of CVD: a pilot 2 x 2 factorial understanding risk trial. Cardiovascular Diabetology. 2008;7:21.

197. Price HC, Griffin SJ, Holman RR. Impact of personalized cardiovascular disease risk estimates on physical activity-a randomized controlled trial. Diabetic Medicine. 2011;28(3):363-72.

198. Price HC, Dudley C, Barrow B, Griffin SJ, Holman RR. Perceptions of heart attack risk amongst individuals with diabetes. Primary Care Diabetes. 2009;3(4):239-44.

199. Price HC, Dudley C, Barrow B, Kennedy I, Griffin SJ, Holman RR. Use of focus groups to develop methods to communicate cardiovascular disease risk and potential for risk reduction to people with type 2 diabetes. Family Practice. 2009;26(5):351-8. Epub 2009/06/24.

200. Karmali KN, Lloyd-Jones DM. Adding a life-course perspective to cardiovascular-risk communication. Nature Reviews Cardiology. 2013;10(2):111-5.

201. van der Weijden T, Bos LB, Koelewijn-van Loon MS. Primary care patients' recognition of their own risk for cardiovascular disease: implications for risk communication in practice. Current Opinion in Cardiology. 2008;23(5):471-6.

202. Roach P, Marrero D. A critical dialogue: communicating with type 2 diabetes patients about cardiovascular risk. Vascular Health and Risk Management. 2005;1(4):301-7.

203. Soureti A, Hurling R, Murray P, van Mechelen W, Cobain M. Evaluation of a cardiovascular disease risk assessment tool for the promotion of healthier lifestyles. European Journal of Cardiovascular Prevention and Rehabilitation. 2010;17(5):519-23.

204. Waldron CA, van der Weijden T, Ludt S, Gallacher J, Elwyn G. What are effective strategies to communicate cardiovascular risk information to patients? A systematic review. Patient Education and Counseling. 2011;82(2):169-81.

205. Visschers VHM, Meertens RM, Passchier WWF, de Vries NNK. Probability Information in Risk Communication: A Review of the Research Literature. Risk Analysis. 2009;29(2):267-87.

206. Hill APF, G.K. Promoting Continuity of Care in General Practice. Royal College of General Practitioners; 2011.

207. Uren Z, Goldring S. Migration trends at older ages in England and Wales. Population Trends. 2007(130):31-40.

208. Royston P, Sauerbrei W. Multivariable model-building : a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester, England ; Hoboken, NJ: John Wiley; 2008.

209. Gatling W, Guzder RN, Turnbull JC, Budd S, Mullee MA, Poole Diabetes S. The Poole Diabetes Study: how many cases of Type 2 diabetes are diagnosed each year during normal health care in a defined community? Diabetes Research and Clinical Practice. 2001;53(2):107-12.

210. Group DAIS. The prevalence of coronary heart disease in Type 2 diabetic patients in Italy: the DAI study. Diabetic Medicine. 2004;21(7):738-45.

211. Uusitupa M, Siitonen O, Aro A, Pyorala K. Prevalence of coronary heart disease, left ventricular failure and hypertension in middle-aged, newly diagnosed type 2 (non-insulin-dependent) diabetic subjects. Diabetologia. 1985;28(1):22-7.

212. Ruigomez A, Garcia Rodriguez LA. Presence of diabetes related complication at the time of NIDDM diagnosis: an important prognostic factor. European Journal of Epidemiology. 1998;14(5):439-45.

213. Cathelineau G, de Champvallins M, Bouallouche A, Lesobre B. Management of newly diagnosed non-insulin-dependent diabetes mellitus in the primary care setting: effects of 2 years of gliclazide treatment--the Diadem Study. Metabolism: Clinical and Experimental. 1997;46(12 Suppl 1):31-4.

214. Baskar V, Venugopal H, Holland MR, Singh BM. Clinical utility of estimated glomerular filtration rates in predicting renal risk in a district diabetes population. Diabetic Medicine. 2006;23(10):1057-60.

215. Knobler H, Zornitzki T, Vered S, Oettinger M, Levy R, Caspi A, et al. Reduced glomerular filtration rate in asymptomatic diabetic patients: predictor of increased risk for cardiac events independent of albuminuria. Journal of the American College of Cardiology. 2004;44(11):2142-8.

216. Rothwell PM, Coull AJ, Giles MF, Howard SC, Silver LE, Bull LM, et al. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). Lancet. 2004;363(9425):1925-33.

217. Eriksen BO, Tomtum J, Ingebretsen OC. Predictors of declining glomerular filtration rate in a population-based chronic kidney disease cohort. Nephron Clinical Practice. 2010;115(1):c41-50.

218. Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, et al. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. American Journal of Epidemiology. 1999;150(4):341-53.

219. Marshall T. Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values. BMC Public Health. 2008;8:25.

220. Clinrisk Ltd. QRISK risk calculator. Available from: http://www.qrisk.org/. [Accessed: 13/6/2014]

221. Carpenter JRK, M. G. Missing data in randomised controlled trials - a practical guide. http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk.pdf. [Accessed: 10/7/2014]

222. Welsh CB, K.; Petersen, I. Application of multiple imputation using the two-fold fully conditional specifcation algorithm in longitudinal clinical data. The Stata Journal. 2014;14(2): 418-431.

223. Medical Research Council. Missing data imputation in clinical databases: development of a longitudinal model for cardiovascular risk factors. Available from: http://gtr.rcuk.ac.uk/project/51521942-5028-44AC-AC17-F3D2929D1BB1. [Accessed: 10/7/2014]

224. Excellence NIfHaC. Obesity: Guidance on the prevention, identification, assessment and management of overweight and obesity in adults and children (CG43). Available from: http://publications.nice.org.uk/obesity-cg43. [Accessed: 10/7/2014]

225. Genuth S, Eastman R, Kahn R, Klein R, Lachin J, Lebovitz H, et al. Implications of the United kingdom prospective diabetes study. Diabetes Care. 2003;26 Suppl 1:S28-32.

226. Wright W, Dent T. Quality standards in risk prediction. Available from: http://www.phgfoundation.org/reports/8685/. [Accessed: 10/7/2014]

227. Pletcher MJ, Pignone M. Evaluating the clinical utility of a biomarker: a review of methods for estimating health impact. Circulation. 2011;123(10):1116-24.

228. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793.

229. Chamnan P, Simmons RK, Khaw KT, Wareham NJ, Griffin SJ. Estimating the population impact of screening strategies for identifying and treating people at high risk of cardiovascular disease: modelling study. BMJ. 2010;340:c1693.

230. DESMOND (Diabetes Education and Self Management for Ongoing and Newly Diagnosed) website. Available from: http://www.desmond-project.org.uk/. [Accessed: 17/7/2013]

231. Visschers VH, Meertens RM, Passchier WW, de Vries NN. Probability information in risk communication: a review of the research literature. Risk Analysis. 2009;29(2):267-87.

232. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-90.

233. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.

234. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ. 2012;344:e4181.

235. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Medicine. 2013;10(2):e1001381.

236. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. Circulation. 2010;121(15):1768-77.

237. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.

238. InPractice Systems. CVD/Stroke Risk Calculators within Vision,2012. InPractice Systems; 2012.

239. Jackson R, Kerr A, Wells S. Vascular risk calculators: essential but flawed clinical tools? Circulation. 2013;127(19):1929-31.

240. Calvert M, Shankar A, McManus RJ, Lester H, Freemantle N. Effect of the quality and outcomes framework on diabetes care in the United Kingdom: retrospective cohort study. BMJ. 2009;338:b1870.

241. Welch CA, Petersen I, Bartlett JW, White IR, Marston L, Morris RW, et al. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. Statistics in Medicine. 2014.