Risk Factors for the Rate of Progression of Chronic Kidney Disease in Secondary Care Patients

Alison Claire Hale
MPhys(Hons), PhD

Lancaster Medical School
Lancaster University
United Kingdom

Thesis submitted for the degree of Master of Science
Lancaster University
November 2019

This thesis is the work of Alison Claire Hale, and has not been submitted for the award of a higher degree elsewhere.

Risk Factors for the Rate of Progression of Chronic Kidney Disease in Secondary Care Patients

Alison Claire Hale

MPhys(Hons), PhD

Thesis submitted for the degree of Master of Science

November 2019

Abstract

Chronic Kidney Disease (CKD) is a major global public health problem and is one of the fastest rising major causes of death. Worldwide moderate to severe CKD has a prevalence of ~11%, whereas in the UK it is ~5%. The objective of our study was to identify key risk factors associated with the progression of kidney disease both across and within primary kidney diseases; ultimately this could lead to improvements in patient care and a reduction in disease burden.

We used data collected from secondary care patients who were recruited into the Salford Kidney Study at Salford Royal NHS Foundation Trust, UK. This ongoing study which commenced in 2002 is one of the largest of its kind worldwide, and consists of over 3000 non-dialysis patients with moderate to severe CKD, who are followed-up annually until an end point of either dialysis, kidney transplant or death. The data recorded at follow-up appointments included comorbidities, medications, lifestyle factors, socio-demographic information and biochemical marker measurements.

We used longitudinal modelling, specifically a linear mixed effects model which models population effects alongside patient-specific variability. We identified risk factors within each of eight primary disease categories including diabetic nephropathy, glomerulonephritis, hypertensive kidney disease, renovascular disease, polycystic kidney disease and pyelonephritis. The key risk factors for lower levels of eGFR are biochemical markers and medications, whereas lifestyle factors and physical attributes are less important. Medications play an important role; in particular ACE inhibitors and ARBs are key in diabetic nephropathy and glomerulonephritis, but not in the other diseases. We found that more rapid progression of kidney disease is associated with biochemical markers including cholesterol and proteinuria. In contrast, medications and comorbidities are not key in rapid disease progression. We recommend future work should include more in-depth studies of each disease category including splitting them into subcategories.

Word count approximately 31,000.

Acknowledgements

I would like to express my thanks to all those who have inspired, challenged and helped me throughout my time at Lancaster University.

Special thanks must go to my main supervisor, Dr Frank Dondelinger, for all his help and support throughout all aspects of the statistical modelling. I would also like to thank Professor Peter Diggle and Professor Philip Kalra who have been prepared to spend time guiding my research by offering helpful and insightful comments.

Additionally I am very grateful to the Medical Research Council for supporting this research through a Skills Development Fellowship.

Contents

List of tables

List of figures

Abbreviations

ACE - angiotensin-converting-enzyme (inhibitor)

AIC - Akaike's information criterion

AKI - acute kidney injury

ARB - angiotensin II receptor blocker

BIC - Bayesian information criterion

BMI - body mass index

CAR1 - first order continuous-time autoregressive (model)

CC - corrected calcium

CCB - calcium channel blocker

CHO - total cholesterol

CI - confidence interval

CKD - chronic kidney disease

CO2 - total (blood) carbon dioxide

cor - correlation

Cr - creatinine

CRIC - Chronic Renal Insufficiency Cohort (study)

CRP - c-reactive protein

CS - compound symmetry

CVD - cardiovascular disease

df - degrees of freedom

DBP - diastolic blood pressure

DN - diabetic nephropathy

EDTA - EthyleneDiamineTetraAcetic acid (anticoagulant)

eGFR - estimated glomerular filtration rate

EPO - erythropoietin (treatment)

FBC - full blood count

GEE - generalised estimating equation (method)

GFR - glomerular filtration rate

GN - glomerulonephritis

Hb - haemoglobin

HbA1c - haemoglobin A1c

HKD - hypertensive kidney disease

HT - hypertension

IHD - ischemic heart disease

IQR - interquartile range

LFT - liver function test

LME - Linear mixed-effects (model)

MDRD - Modification of Diet in Renal Disease (Study)

MICE - Multiple Imputation by Chained Equations

NHS - National Health Service (UK)

NICE - National Institute for Health and Care Excellence (UK)

ONS - Office for National Statistics (UK)

PKD - polycystic kidney disease

PN - pyelonephritis

PO - phosphate

PP - systemic pulse pressure (systolic minus diastolic blood pressure)

PTH - parathyroid hormone

Pu - proteinuria

RIP - rest in peace - relates to patients who died while part of the study

RRT - renal replacement therapy; that is haemodialysis dialysis, peritoneal dialysis or kidney transplant

RVD - renovascular disease

SBP - systolic blood pressure

sd - standard deviation

se - standard error

SKS - Salford Kidney Study (UK)

SRFT - Salford Royal NHS Foundation Trust (UK)

U&E - urea and electrolytes

UK - United Kingdom

USRDS - United States Renal Data System

VIF - variance inflation factor

var - variance

Mathematical notation

$\forall$ - for all

i - patient (subject) index

j - time index

M - number of patients

n_i - number of observations for patient i

$\dot{x}(t)$ - dot denotes time derivative of variable $x(t)$ i.e. $\frac{d}{dt}(x(t))$

$\hat{\beta}$ - estimate of β

$\boldsymbol{I}$ - identity matrix

$\mathbf{1}$ - matrix of ones

$\boldsymbol{X}^T$ - T denotes transpose of matrix $\boldsymbol{X}$

$X \sim \mathrm{N}(\mu, \sigma^2)$ - random variable X distributed normally with mean μ and variance σ^2

$\mathbb{E}(X)$ - expectation of random variable X

$P(A|B)$ - conditional probability of A given B

$\mathbb{S}_i(X_t)$ - cubic spline interpolation over time of X_{ij} at time points t_{ij} for patient i

$\mathrm{cor}(\cdot, \cdot)$ - correlation function

$\mathrm{var}(\cdot)$ - variance function

$|x|$ - absolute value of x

$x^{'}$ - superscript dash denotes x belongs to the standardised model. Note dash does not denote the derivative of x

Variable definitions

- The reference level for each categorical variable is denoted using italic font. For example if the levels are '*non-smoker*', 'active' and 'ex-smoker' then the reference level is non-smoker.
- Variables measured only at baseline have names which end in a zero e.g. baseline age is denoted 'age0'.
- Variables measured at baseline and also subsequent follow-up appointments omit the trailing zero in their name e.g. 'age'.
- An *interaction* between each time varying covariate and time since baseline (followupTime) is denoted using a colon, e.g. the interaction between Hb and follow-up time is written Hb **:** $followupTime$.

age0 - age at baseline appointment in units of years

age - age in units of years at follow-up appointment

bodyMassIndex - body mass index

CC - corrected calcium

comorbidityCancer - denotes if the patient has/had any type of cancer - levels '*no*', 'current', 'previous'

ComorbidityCV - number of cardiovascular conditions the patient has - levels '*no*', '1', 'over 1'; more details Appendix A.1.3

comorbidityDiabetes - denotes if patient has diabetes - levels '*no*', 'type 1', 'type 2'; note for disease 'diabetic nephropathy' the reference level is '*type 2*' as all patients have diabetes

comorbidityGastrointestinal - denotes if patient has any long-term gastrointestinal disease(s) - levels '*no*', 'yes'; more details in Appendix A.1.3

comorbidityOther - denotes if patient has any long-term conditions not included in the above categories - levels '*no*', 'yes'; more details in Appendix A.1.3

Cr - creatinine

CRP - c-reactive protein

DBP - diastolic blood pressure

disease - primary kidney disease of each patient - levels '*other*', 'diabetic nephropathy', 'glomerulonephritis', 'hypertensive kidney disease', 'obstruction', 'polycystic kidney disease', 'pyelonephritis', 'renovascular disease', 'unknown'

endDate - date when the patient leaves the study which is always due to either dialysis, kidney transplant or death (whichever happens first)

endReason - reason for patient leaving study - levels '*ONGOING*', 'LOST', 'PRESUME_LOST', 'RIP', 'RRT' - more details in Appendix A.1.1

ethnicity - patients are categorised as either '*White*' or 'nonWhite'

familyHistoryIHD0 - family history of ischemic heart disease - recorded at baseline - levels '*no*', 'yes'

followup - integer number of years between baseline appointment and a given follow-up appointment

followupTime - time interval between baseline appointment and a given follow-up appointment - this is a real number with units of years

Hb - haemoglobin

HbA1c - haemoglobin A1c

livingStatus0 - whether or not the patient is living alone - only recorded at baseline - levels '*with others*', 'alone'

logeGFR - natural logarithm of the estimated glomerular filtration rate

med.ACE.ARB - patient is on ACE inhibitor and/or ARB medication - levels '*no*', 'yes' - details in Appendix A.1.4

med.AlphaBlockers - patient is on alpha blocker medication - levels '*no*', 'yes' - details in Appendix A.1.4

med.BetaBlockers - patient is on beta blocker medication - levels '*no*', 'yes' - details in Appendix A.1.4

med.CCBs - patient is on calcium channel blocker medication - levels '*no*', 'yes' - details in Appendix A.1.4

med.Diuretics - patient is taking a diuretic - levels '*no*', 'yes' - details in Appendix A.1.4

med.Epo - patient had at least one erythropoietin treatment since their last follow-up - levels '*no*', 'yes'

med.Iron - patient is taking an oral iron supplement (N.B. does not include iron injections) - levels '*no*', 'yes'

med.Other - patient is on medication which does not come under one of the other **med.xxx** categories - levels '*no*', 'yes' - details in Appendix A.1.4

med.ParenteralIron - patient has been administered iron injections since their previous follow-up (N.B. does not include iron taken orally) - levels '*no*', 'yes'

med.VitaminD - patient is taking a vitamin D supplement - levels '*no*', 'yes' - details in Appendix A.1.4

numberAKIepisodes - number of AKI episodes since last follow-up appointment

numberAntihypertensives - count of distinct antihypertensives drugs the patient is taking at each follow-up appointment

numberClinicVisits - number of visits to the renal clinic since last follow-up appointment

occupation0 - patient's occupation - only recorded at baseline - levels '*RoutineManual*', 'Managerial-Professional', 'Intermediate' , 'NeverWorkedUnemployed' - more details in Appendix A.1

PO - phosphate

PP - systemic pulse pressure (systolic minus diastolic blood pressure)

PTH - parathyroid hormone

Pu - proteinuria

SBP - systolic blood pressure

sex - patient's sex - levels '*male*', 'female'

smokingStatus0 - patient's smoking status - only recorded at baseline - levels '*non-smoker*', 'active', 'ex-smoker'

StudyID - each patient's unique identifier in the SKS

totalCholesterol - total Cholesterol

totalCO2 - total CO_2

weeklyAlcohol0 - number of units of alcohol the patient typically consumes within a week - only recorded at baseline - levels '*under 1*', '1 to 14', 'over 14'

1 Introduction and background

CKD is recognised as a major global public health problem with a high economic cost to health systems (1). The 2015 Global Burden of Disease Study (2) reported kidney disease as the 12th most common cause of death, with CKD mortality increasing by 31.7% between 2005-2015, it is now one of the fastest rising major causes of death worldwide (3). This growth is generally considered to be fuelled by overnutrition, inadequate physical inactivity, and ageing populations (4,5). More broadly the World Health Organization confirms a global shift in which the majority of global morbidity and mortality is now caused by chronic diseases as opposed to infectious diseases (6,7). For moderate to severe CKD, stages 3 to 5, the global prevalence was reported in 2016 to be 10.6% {95% CI: 9.2-12.2%}; see (8). In 2014 Public Health England estimates, which took account of both diagnosed and undiagnosed cases, indicated a prevalence of 6.1% {95% CI: 5.3-7.0%} for adults with CKD stages 3 to 5 who were resident in England (9). This rate is similar to the actual diagnosed prevalence of 4.3% reported by the Quality and Outcomes Framework during 2012-2013; see (10,11). The prevalence of CKD dramatically increases with advancing age (12). For example, (13) reported in 2007 that the prevalence in the United States of CKD stage 3 stratified by age was: 20-39 years (~1%); 40-59 years (~4%); 60-69 years (~14%); $\geqslant$ 70 years (~37%). This study also showed that stage 3 was by far the most prevalent out of all the five stages of CKD.

CKD is generally associated with decreased quality of life along with an increased risk of premature death and cardiovascular disease (14). It follows that a rapid decline in kidney function is associated with an increased risk of both mortality and cardiovascular events (15,16). Conversely, cardiovascular disease increases the risk of CKD hence these two diseases are closely interrelated (17). CKD is also frequently comorbid with other common diseases including hypertension, diabetes, anaemia and mineral/bone disorders (18,19), in fact diabetes and hypertension are the leading causes of CKD (20,21). For example, during 2017, the United States Renal Data System (USRDS) reported (in chapter 1) that given adults with CKD (stages 1-5), about 40% had diabetes, ~32% had hypertension and ~42% had cardiovascular disease (18). The prevalence of comorbidities increases as CKD progresses and a majority of patients with moderate to severe CKD have at least one comorbidity (22). The primary causes of end-stage renal disease, as reported by USRDS, are diabetes 38.2%, hypertension 25.5% and glomerulonephritis 16%; see table 1.6 in (23). Mortality rates are also substantially higher for certain groups of CKD patients. In particular the mortality rate for CKD patients with cardiovascular disease is about 2.5 times higher than for those without cardiovascular disease or diabetes, similarly the mortality rate for CKD patients with both cardiovascular disease and diabetes is about 3 times higher than for those without cardiovascular disease or diabetes; see (23) chapter 3. Given that for CKD patients the risk of complications increases with decreasing kidney function, early intervention aims to ameliorate the risk of severe complications and reduce the number of patients progressing to

dialysis or transplant e.g. see (24–26).

To determine how well the kidneys are functioning the level of creatinine in the blood is measured. This measured value is then used to calculate the estimated glomerular filtration rate (eGFR). Normal kidney function in healthy adults decreases with age; for example adults of 20-30 years have an eGFR of ~115 mL/min/1.73m^2 whereas it has decreased to ~85 mL/min/1.73m^2 in the 60-69 year age group (27,28). The annual rate of decline of eGFR in the healthy population is approximately 0.36-1.21 mL/min/1.73m^2 per year; younger adults tend towards the lower value and older individuals the upper value; see reviews (28) and (29). It should be noted that in the general population the aforementioned values vary widely as they not only depend on factors such age, ethnicity, gender but are also dependent on underlying comorbidities. The National Institute for Health and Care Excellence (NICE), defines progressive CKD as either an annual fall in eGFR of $\geqslant$ 5 mL/min/1.73m^2 or a fall of $\geqslant$ 10 mL/min/1.73m^2 within 5 years (30). Furthermore it is generally accepted, as defined by KDIGO in 2012, that rapid progression is a sustained decline of $\geqslant$ 5 mL/min/1.73m^2 per year (31). CKD can be divided into several primary disease types including glomerulonephritis, diabetic nephropathy and polycystic kidney disease. These diseases are expected to have different rates of decline in eGFR although exact values vary widely in the literature and are often not directly comparable. However in 2012/13, (32) reported an average annual decrease for diabetic nephropathy patients of 1.7 mL/min/1.73m^2 whereas (33) found an average annual decrease of about 3 mL/min/1.73m^2 in polycystic kidney disease patients. This suggests that the progression of CKD is nearly twice as fast in polycystic patients; both rates were for patients with CKD stages 3 to 5.

In this thesis we study the progression of CKD using data collected by the ongoing Salford Kidney Study (SKS) (34,35) run by Salford Royal NHS Foundation Trust (SRFT), UK. SKS has one of the largest cohorts in the world of secondary care CKD patients, with over 3000 patient records collected since 2002. The data includes patients with all primary kidney disease types. The aims of the SKS are to investigate factors influencing outcomes and progression of renal disease in CKD patients, including a focus on risk factors associated with more rapid disease progression. In particular, SKS is a prospective observational study of outcomes of non-dialysis adult patients with CKD stages 3 to 5 ($10 <$ eGFR $\leqslant 60$ mL/min/1.73m^2). Patients referred to the renal services at SRFT, and existing CKD patients attending the clinics, are approached for inclusion in the study and enrolled if written informed consent is obtained. Patients are followed up annually until they reached predefined study end-points, these are death or initiation of renal replacement therapy (RRT). SKS defined RRT as chronic haemodialysis, peritoneal dialysis or kidney transplant. At recruitment and annual nephrology follow-up appointments, patient socio-demographic and lifestyle choices are recorded along with comorbidities. Concurrent medications and additionally blood samples are taken and processed to obtain a comprehensive set of biochemical marker measurements.

In general, longitudinal data such as the SKS data, is comprised of multiple observations collected over successive time periods on the same individuals. The data may also include baseline variables that are collected once e.g. age at study entry. However repeated measurements on the same individual will not be independent and this must be accounted for when building statistical models. To this end mixed effects models are an appropriate statistical framework and a well-established approach; for example see textbooks (36–38). These models consist of both *fixed effects* and *random effects*, which explain the relationships between an outcome variable and explanatory variables. Fixed effects describe the whole population whereas random effects are associated with each individual and capture the dependence of repeated measurements. In terms of longitudinal data the development of such models is attributed to Laird and Ware in 1982 (39); this paper considers a causal link between air pollution and pulmonary function measured at specified time intervals. Later in 1988 Diggle (40) introduced an approach whereby the correlation between successive random effects is described by stationary Gaussian processes; this approach is applied to two separate repeated measure studies, body weight of rats and blood pressure of rats.

Mixed effects models have been extensively used to study the progression of kidney disease over time. A broad literature review of statistical methods used for investigating risk factors of CKD progression is given by (41). One of their conclusions, given longitudinal data where the outcome of interest is the entire trajectory of renal function over time, is that linear mixed models are an appropriate tool for estimating both risk factors and their associated confidence intervals. Given a choice between linear regression to estimate individual slopes and linear mixed effects models, (42) concludes the latter are preferred for research questions regarding kidney disease trajectories over time at population level. Similarly in the context of progression of kidney disease (43) considers the comparative strengths and weaknesses of the Generalized Estimating Equations (GEE) approach with linear mixed effects models, in part concluding that the mixed effects model is preferred in relation to missing data since the GEE makes more restricted assumptions; for details see Appendix 4 in the supplementary material of (44). A further comparative study by (45) concludes that the linear mixed model is the preferred method for investigating risk factors associated with renal function trajectories when individuals leave the study due to initiation of renal replacement therapy.

In this thesis, we performed a longitudinal analysis of the SKS data, to identify markers for progression in CKD. The patients were assigned to one of 8 subcategories of CKD, we refer to these as primary disease categories. We applied a linear mixed model (LME) to analyse each of the 8 primary disease categories separately, and used model selection techniques to identify the most pertinent risk factors. As a result we were are able to make comparisons across the primary disease categories.

We start, in Chapter 2, by exploring and summarising the SKS data. In Chapter 3 we define the LME which forms the basis of all our modelling. In Chapter 4 we show how to interpret step

changes in the LME model regression parameters in terms of eGFR (rather than log(eGFR)) and also how to use to estimate the rate of change over time of eGFR from the LME model. We describe our model selection procedures in Chapter 5 and then having selected the final model for each primary disease category we then validate each model using diagnostic procedures before presenting our results in Chapter 6. Our findings are reported in Chapter 7. In Chapter 8 we discuss our models, results and future research directions. We close, in Chapter 9, with some concluding remarks.

2 Summary of SKS data

We begin by describing our procedures for cleaning the raw SKS data. This includes removing obvious erroneous values and consolidating subsets of data into categories such as primary diseases, comorbidities and medications. The cleaned dataset has approximately 40 potential risk factors (explanatory variables) which we use during our exploratory analysis. Finally, after completing the exploratory analysis, the number of complete records was significantly increased by imputing missing values thereby increasing the power of our statistical models. Throughout this chapter, unless otherwise stated, missing values are not imputed.

2.1 Data preparation and cleaning

Using the programming language R (46) we extracted and cleaned the SKS data from the Microsoft Access database provided by the clinicians at Salford Royal NHS Foundation Trust. All incorrect data were purged, for example a date with year 1066. The units of all measurements were converted so as to be consistent e.g. patient heights were standardised to metres. We accounted for spelling variations and commonly misspelt words e.g. medications 'doxazosin' and 'doxasosin' were both identified as α-blockers. To reduce the complexity of the data we, with guidance from the clinicians, categorised various items; notably medications, comorbidities and primary kidney diseases. The breakdown of these categories is given in Appendix A.1. The biochemical marker data was provided separately from the Microsoft Access database, so where possible we matched the biochemical data to each patient using their follow-up appointment dates; we allowed for differences of up to six weeks between the recorded dates of the biochemical markers and follow-up appointments. Full details regarding data cleaning are given in Appendix A.1.

2.2 Overview of SKS data

The data from 3,166 patients were collected between 01 October 2002 and 27 February 2017; participants were recruited throughout this period. Of the patients in this study 37.6% were female, and 95.7% declared their ethnicity as white.

At baseline, when the patient joined the study, a number of health indicators were recorded. For example the cohort had 12.2% active smokers and 52.7% ex-smokers. Similarly within the cohort 29.9% of patients declared they consumed 1 to 14 units of alcohol per week while another 14.7% declared they drank over 14 units per week. Further basic summary statistics of the cohort at baseline are given in Table 1; note IQR refers to interquartile range. These show that the cohort are on average older adults who are, as defined by NICE, overweight (47). Within the general UK population pulse pressure (PP) for adults aged around 65 years is expected to be

in the upper fifties (48) so the SKS cohort is a little worse than average but 87.1% are taking antihypertensives.

Table 1: Baseline summary statistics

item	units	min	max	median	IQR
age	year	18.2	94.5	67.4	20.0
BMI	kg/m^2	13.3	59.9	28.0	7.8
DBP	mmHg	40.5	137.0	74.5	14.0
PP	mmHg	17.0	146.0	64.0	28.0
SBP	mmHg	76.0	218.0	139.0	29.0

Given all patients, including those who have not reached an end point, the average time in the study was 4.6 years, with 7 patients reaching 14 annual follow-up years. There were 606 patients who left the study to undergo renal replacement therapy (RRT); in the SKS RRT is defined as haemodialysis dialysis, peritoneal dialysis or kidney transplant. In addition 952 patients died while part of the study, and 99 patients who were lost to follow-up. The average time patients were in the study before RRT or death was 3.9 years. Of the remaining 1313 patients in the study there were 699 with a time span of more than 2 years 6 months since their last follow-up appointment.

2.3 Primary kidney disease types

We categorised the patients as having one of the following primary kidney diseases: diabetic nephropathy, glomerulonephritis, hypertensive kidney disease, obstruction, other, polycystic kidney disease, pyelonephritis, renovascular disease, unknown. Figure 1 shows their frequencies.

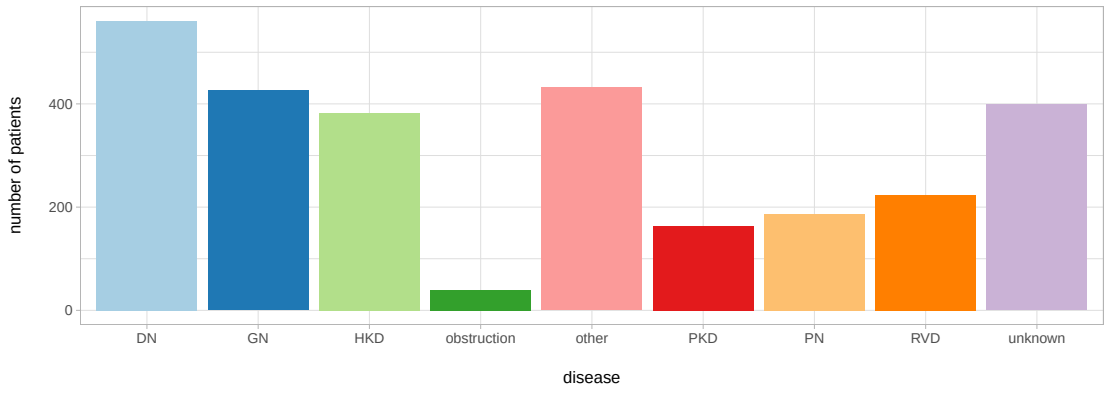

Figure 1: Primary disease type frequency

See the Appendix A.1.2 for the clinical breakdown of conditions/diseases within each primary disease category. The basic characteristics of these diseases are:

- *diabetic nephropathy* (DN) - chronic loss of kidney function occurring in patients with diabetes.
- *glomerulonephritis* (GN) - refers to several kidney diseases many of which are characterised by inflammation within specific kidney sub-structures.
- *hypertensive kidney disease* (HKD) - chronic high blood pressure causes damage to the kidney tissue. Usually these patients do not have a renal biopsy.
- *obstruction* - obstructive nephropathy - has a number of causes but is characterised by a blockage in the flow of urine out of the kidney(s).
- *polycystic kidney disease* (PKD) - is a genetic disorder causing the growth of multiple cysts within the kidneys.
- *pyelonephritis* (PN) - inflammation of the kidney often caused by a bacterial infection.
- *renovascular disease* (RVD) - has a number of causes and is characterised by a progressive narrowing or blockage of the large renal arteries or veins.
- *other* - all other primary kidney diseases which are less common and as such they do not fall into the aforementioned disease categories.
- *unknown* - refers to chronic renal failure when the aetiology is uncertain, unknown or unavailable. This is a heterogeneous disease grouping whose common characteristic is that the patient's kidney disease is not clinically identified. For example given a patient with exceptionally slow disease progression it may be unjustified to do an invasive procedure such as a biopsy to confirm the cause of their disease.

2.4 Comorbidities

Comorbidities were recorded at baseline and thereafter at each follow-up. We collated comorbidities into the following clinically relevant categories where percentages indicate the proportion of patients recorded as having a given comorbidity at some point while in the study:

- 78.2% cardiovascular disease
- 35.4% diabetes
- 25.4% other
- 10.1% gastrointestinal disease
- 3.8% had cancer during the study. We note 16.3% had cancer either during the study or at a previous time.

Under this classification 54.8% of patients have multiple comorbidities. The cancer, cardiovascular and diabetes categories can be subdivided into specific diseases, for example of the patients with diabetes 87.2% had type 2. Appendix A.1.3 gives details of the conditions/diseases which are

included in each comorbidity category.

2.5 Medications

Medication and treatment data were also recorded at baseline and thereafter at annual follow-up appointments. At baseline 87.1% were taking at least one antihypertensive. Here medications are grouped as follows where percentages indicate the proportion of patients taking a given medication at some time during the study:

- 69.2% angiotensin-converting-enzyme (ACE) inhibitor and/or angiotensin II receptor blocker (ARB)
- 58.9% diuretic
- 54.3% calcium channel blocker (CCB)
- 42.1% β-blocker
- 38.6% α-blocker
- 32.4% vitamin D
- 27.4% EPO treatment (for anaemia)
- 24.4% iron taken orally
- 23.1% iron administered by injection

In addition we noted that 68.9% were on statins and 43.7% took aspirin. All other medications not mentioned above occurred less frequently in the data than iron taken orally. Details of the drugs in each category can be found in the Appendix A.1.4.

2.6 Biochemical markers

2.6.1 General biomarkers

In addition the study also measured biochemical markers from blood and urine samples during annual follow-up appointments and other hospital visits e.g. AKI episodes. Standard laboratory markers from blood samples included: full blood count (FBC), urea and electrolytes (U&E), liver function test (LFT), calcium, phosphate, cholesterol, Parathyroid Hormone (PTH). Furthermore EDTA whole blood, serum, plasma, and citrate plasma samples were processed and stored at -800C. Table 2 lists the biochemicals pertinent to this thesis; except for creatinine they enter into our models as explanatory (input) variables.

Table 2: Summary statistics for biochemical markers at baseline

biochemical	units	min	max	median	IQR
CRP - c-reactive protein	mg/L	0.10	195.0	3.4	6.2
CHO - total cholesterol	mmol/L	2.10	16.0	4.4	1.5
CC - corrected calcium	mmol/L	1.21	3.0	2.3	0.2
Cr - creatinine	μmol/L	51.00	915.0	179.0	126.0
CO2 - total CO2	mmol/L	10.50	44.5	23.0	4.5
Hb - haemoglobin	g/L	10.90	195.0	122.0	24.0
HbA1c - haemoglobin A1c	mmol/mol	25.00	154.0	50.0	24.0
PO - phosphate	mmol/L	0.43	3.2	1.1	0.3
PTH - parathyroid hormone	pmol/L	0.32	99.1	7.1	8.7
Pu - proteinuria	g/24hr	0.02	17.2	0.3	0.9

We assume the variables are independent in our statistical models, Table 3 confirms there is no significant correlation between the biochemicals. The only exception is a strong negative correlation between creatinine and eGFR which is to be expected given the formula for calculating eGFR includes a creatinine term; see Equation 1.

Table 3: Correlation between biochemical markers for all follow-up years

	CC	CHO	CO2	Cr	CRP	eGFR	Hb	HbA1c	PO	PTH	Pu
CC	1	0.1	0.2	-0.2	0.0	0.1	0.0	0.0	-0.1	-0.2	-0.1
CHO		1.0	0.0	-0.2	0.0	0.2	0.1	0.0	-0.1	-0.1	0.2
CO2			1.0	-0.3	-0.1	0.3	0.1	0.1	-0.3	-0.2	-0.2
Cr				1.0	0.1	**-0.8**	-0.3	0.0	**0.6**	0.5	0.2
CRP					1.0	-0.1	-0.2	0.0	0	0.1	0.0
eGFR						1	0.3	0.0	-0.5	-0.4	-0.2
Hb							1.0	0.0	-0.4	-0.2	-0.1
HbA1c								1.0	0.1	0.0	0.1
PO									1	0.4	0.3
PTH										1.0	0.1
Pu											1.0

2.6.2 Estimated glomerular filtration rate (eGFR)

Glomerular Filtration Rate (GFR) is a key indicator of renal function, its estimate eGFR is derived from a patient's serum creatinine level, age, sex and race. Creatinine is a compound produced by metabolism of creatine and is excreted in the urine. In healthy individuals the kidneys maintain blood creatinine in a normal range, an elevated creatinine level indicates impaired kidney function. In our statistical models the outcome variable will be eGFR, our primary motivation for using eGFR as opposed to creatinine is that clinicians advised us that they find eGFR easier to interpret. Hence eGFR is a clinically reasonable indicator of kidney function. Table 4 gives the standard definitions of CKD stages in terms of eGFR (30,31); stage 1 is mild impairment whereas stage 5 signifies kidney failure.

stage	1	2	3	4	5
eGFR	$\geqslant$90	89 - 60	59 - 30	29 - 15	$<$15

Table 4: CKD stage defined by eGFR (mL/min/1.73m^2)

There are several equations for estimating GFR (49) however it is mostly agreed that in general the CKD-EPI equation gives the best estimate (50–52). Additionally given NICE (30) recommends this equation we use it for calculating eGFR in units mL/min/1.73m^2

$$\begin{aligned}\text{eGFR} = 141 \times \min(\text{S}_{\text{cr}}/\kappa, 1)^{\alpha} \times \max(\text{S}_{\text{cr}}/\kappa, 1)^{-1.209} \times 0.993^{\text{age}} \times \\ 1.018[\text{if female}] \times 1.159[\text{if black}]\end{aligned} \tag{1}$$

where

- S_{cr} is serum creatinine with units μmol/L
- κ is 61.9 for females and 79.6 for males
- α is -0.329 for females and -0.411 for males
- $\min(\text{S}_{\text{cr}}/\kappa, 1)$ indicates the minimum of either $\text{S}_{\text{cr}}/\kappa$ or 1
- $\max(\text{S}_{\text{cr}}/\kappa, 1)$ indicates the maximum of either $\text{S}_{\text{cr}}/\kappa$ or 1
- age has units of years

At follow-up appointments we find the median eGFR across all patients is 28.1 with interquartile range (IQR) 23.3. Hence the patient's generally have moderate to severe CKD; stages 3 and 4. In contrast if we consider only acute kidney injury (AKI) episodes the overall median eGFR drops to 14.6 with IQR 17.8.

Given all patients at follow-up, eGFR follows a right skewed distribution; e.g. ($\text{mean}_{\text{eGFR}} = 31.5$) > ($\text{median}_{\text{eGFR}} = 28.1$). Figure 2 is used for exploratory purposes only, the qq-plot in panel

(a) shows the distribution of the log of eGFR to be approximately normal; visual confirmation of the distribution's shape is given by the histogram. Applying the Shapiro–Wilk normality test (53) to the log(eGFR) distribution yields a p-value <0.0001 hence we reject the null hypothesis and conclude it significantly deviates from normality. In our statistical models we choose to use log(eGFR) as the outcome variable. Given log(eGFR) is closer to a normal distribution than eGFR it is expected to give a better empirical fit of our data to the models, for further details see Chapter 3. From Equation 1 we note that log(eGFR) is equivalent to creatinine adjusted for age and sex however in our models we will consider using age and sex as explanatory variables because Equation 1 has been shown not to be optimal for all sub-populations; e.g. (49) and (54). Note that when we write log(eGFR) this denotes $\log_e(e_0^{-1}\text{eGFR})$ where constant e_0 equals 1 and carries the same physical dimensions (units) as eGFR, this ensures the argument of the logarithm does not have physical dimensions.

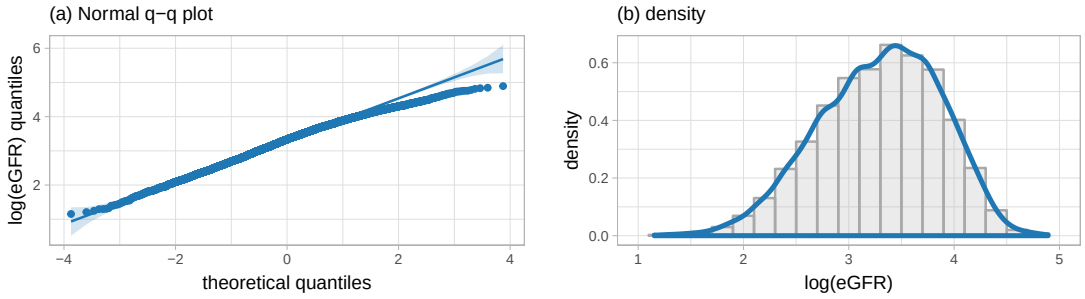

Figure 2: Distribution of log(eGFR) for all patients at follow-up

Considering all log(eGFR) values from a random selection of patients, in Figure 3 we see that the progression of CKD over time is far from a smooth monotonic function. However these figures include measurements taken between follow-up appointments when the patients will in some cases be experiencing an acute episode of illness e.g. AKI. Grouped by disease Appendix A.2, Figures 54 to 62, depicts Trellis plots for an arbitrary selection of patients showing the log of their eGFR at each follow-up year; these figures show although there is much individual variation most patients have an approximately linear downward trend in log(eGFR) as time passes.

Given each primary kidney disease, Figure 4 (a) shows log(eGFR) values for every patient at each follow-up, where red points are the marginal means at each follow-up time. Figure 4 (b) depicts the corresponding variances. We note that both the mean and variance are less informative when there are fewer observations for example in later follow-up years. We observe, in Figure 4 (a), that successive marginal means (red points) for most disease categories exhibit an overall downward trend as the number of follow-up years increase. If we naively ignore the correlation between observations on the same individual and fit straight lines through the marginal mean points for each disease we find, for instance, that on average PKD patients loose kidney function 1.8 times faster than those with diabetic nephropathy.

In Chapter 3 we will use rigorous statistical modelling to explore the progression of disease while accounting for the explanatory variables discussed above e.g. demographics, comorbities, medications, etc.

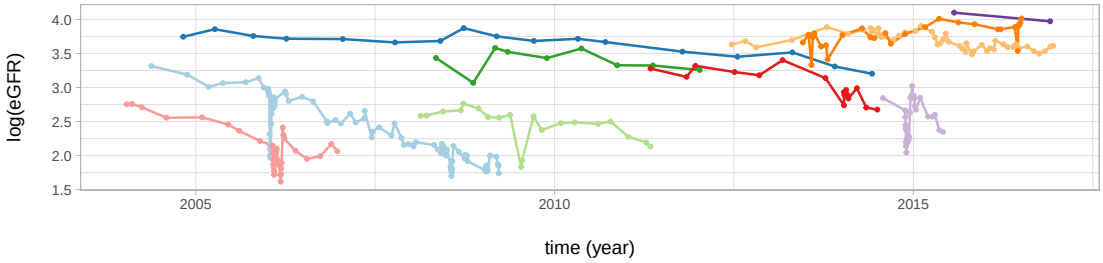

Figure 3: eGFR progression of an arbitrary sample of 10 patients

2.7 Imputation

Prior to this section we have not imputed missing values. In our cleaned version of the SKS dataset we assumed all missing data values were missing completely at random unless there was evidence to the contrary. In particular we assumed each missing value was: independent of the values of other variables (fields); independent of the value of the observation; and independent of time. The proportion of missing values in our cleaned dataset was 7.4%. This level of missing values diminishes the potential statistical power of our models. Therefore to improve statistical power imputation methods were employed. Appendix A.3 lists all continuous and categorical variables for which missing values were imputed.

Popular imputation methods include Multiple Imputation (55,56) and Expectation-Maximization (57) of which there are many extensions and algorithms, two examples respectively are Multivariate Imputation by Chained Equations (58) and Amelia (59). All such methods are intended for multivariate data and rely on correlations between variables (inter-variable) to estimate missing values. In our case we treat each variable (field) for a given patient as a timeseries consequently these methods cannot be directly applied because a timeseries is univariate and exhibits inter-time (intra-variable) correlations; for example see (60) for an overview of timeseries imputation methods. In this thesis we employ imputation algorithms which are specifically intended for use with timeseries data; in particular we use the R-package imputeTS (61) to impute all missing values.

For continuous variables (e.g. BMI) we use Kalman smoothing on a structural model fitted by maximum likelihood; for example (62,63) give methodological details. By design this imputation method accounts for temporal trends, hence it is appropriate for our data where we often observe trends e.g. a patient's BMI may gradually increases/deceases over several successive follow-up years. All our continuous variables have values which are always positive so to overcome the problem

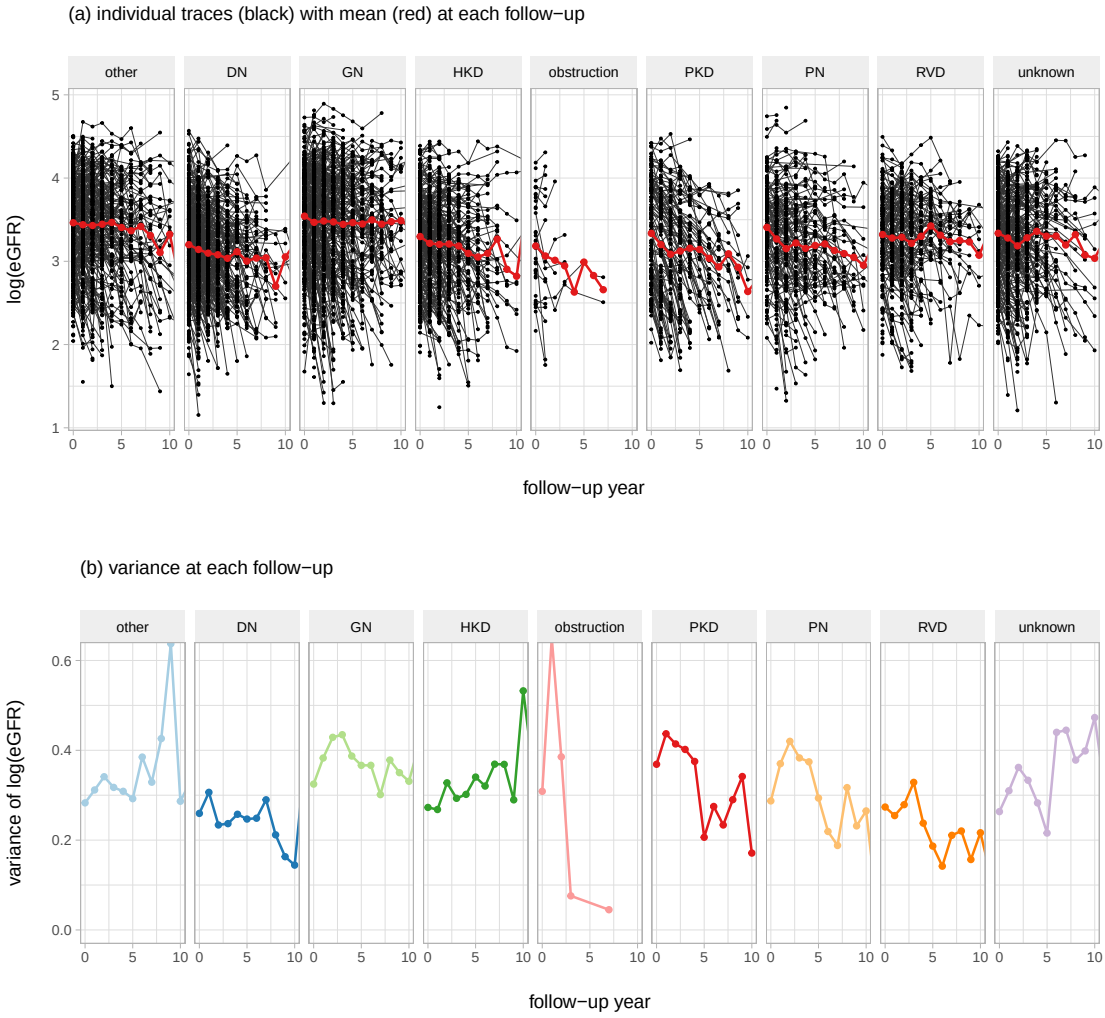

Figure 4: eGFR values of study cohort grouped by disease

of the imputation producing negative values we did the following: (a) use logarithm function to transform the variable onto the logarithm scale; (b) impute missing values; (c) transform the variable back to its original scale with the antilogarithm function. In terms of categorical variables (e.g. weekly alcohol intake) missing values are estimated with Last Observation Carried Forward/Backward methods where priority is given to Forward imputation, in other words where possible the last observed value is carried forward in time to subsequent follow-up appointments.

The SKS data explicitly recorded the existence of a comorbidity but did not explicitly record if it was not present; an empty comorbidity field implied the patient did not have the given comorbidity at a particular follow-up year. Each patient's comorbidities were frequently not, or only partially, recorded at each follow-up. Consequently, the data suggested that many patients recovered from, and were often subsequently re-inflicted with, long-term health conditions such as dementia. Since this is implausible for long-term conditions we assumed that each patient's condition(s) persisted for all future time after the follow-up at which it was first recorded; this approach was applied to all comorbidities listed in Appendix A.1.3. Prior to the first instance of a comorbidity being recorded we assumed that the patient did not have the condition.

At each follow-up *all* the medications for each patient were typically documented; we assume if at least one drug/supplement was recorded then all drugs/supplements were recorded. At a given follow-up, if at least one medication is recorded then we assign the patient as either taking, or not taking, a drug/supplement in each of our medication categories. Conversely, if no medications are recorded we impute using the same approach as we used for comorbidities. This is the reason all medication categories, except for EPO treatment, have the same number of missing records before imputation (and also after imputation); see Table 5. We dealt with both EPO treatment and parenteral iron separately from the other medications as these are not recorded as part of the SKS medication lists. These are administered intermittently so unless recorded we assume the patient did not receive the treatment.

Biochemical measurements were recorded at follow-up appointments but unlike the rest of the SKS data they were also recorded at other hospital/clinic visits. The data recorded outside of follow-up appointments would sometimes relate to episodes of acute illness (e.g. AKI). During acute illness some, or all, of the biochemical measurements could potentially be very different, for example as discussed in Section 2.6.2 the cohort median eGFR is 38% lower during identifiable AKI episodes compared with follow-up appointments. Finding a method to robustly identify all acute episodes is beyond the scope of this thesis. Consequently to impute missing values at follow-up appointments we only used measurements recorded at either past or future follow-ups.

In instances where a patient had no recorded values for a given field (over all their follow-ups) we did not impute values; creating imputation models for these rare instances was beyond the scope of this thesis. If a patient's timeseries had only one recorded value we duplicated this value at all points in the series, we did this for all relevant continuous and categorical variables except

medications and comorbidities.

Table 5 shows the proportion of missing values for each variable before and after imputation, as can be seen imputation substantially reduces the number of missing values. In this table we include the proportion of missing creatinine values because this directly affects, and is the main contributor to, the proportion of missing eGFR values.

As is seen in Table 5 HbA1c has a very high number of missing values; this is because it is generally only recorded in patients with diabetic nephropathy. In this sub-group before imputation the percentage of missing HbA1c is 68.3% and after imputation 30.7%. In the next chapter we will only use HbA1c for models relating to diabetic nephropathy patients, however given the high quantity of missing data it may adversely affect the statistical power of such models and given the large quantity of imputed values it may not be informative; we reserve judgement until we obtain the model results.

Summary statistics for each continuous variable before and after imputation confirm the imputed values did not significantly alter the overall distribution of any continuous variable; see results tabulated in Table 6. For a given patient and follow-up year we define a 'complete record' as having all values for every variable of interest. If HbA1c is omitted, then before imputation there were 2024 complete records and after 3121, therefore the imputation of missing values will substantially increase the statistical power of our models. For the remainder of this thesis we use the cleaned SKS data augmented with imputed values.

Table 5: Proportion of missing values before and after imputation over all follow-up years

group	item	Before (%)	After (%)
general	BMI	16.1	4.3
	DBP	4.0	0.8
	number of antihypertensives	4.9	0.4
	PP - pulse pressure	4.0	0.8
	SBP	3.8	0.8
biochemical	CC - corrected calcium	1.9	0.1
	CHO - total cholesterol	22.2	3.9
	CO2 - total CO2	16.0	2.3
	Cr - creatinine	0.0	—
	CRP - c-reactive protein	30.9	4.2
	eGFR	0.7	—
	Hb - haemoglobin	1.6	0.2
	HbA1c - haemoglobin A1c	87.8	64.9
	PO - phosphate	2.7	0.2
	PTH - parathyroid hormone	20.8	2.5
	Pu - proteinuria	11.4	2.3
categorical	comorbidity cancer	3.8	0.1
	comorbidity cardiovasular	3.9	0.0
	comorbidity diabetes	4.2	0.1
	comorbidity gastrointestinal	4.8	0.0
	comorbidity other	3.8	0.1
	medication ACE and/or ARB	5.3	0.6
	medication alpha blockers	5.3	0.6
	medication beta blockers	5.3	0.6
	medication CCBs	5.3	0.6
	medication diuretics	5.3	0.6
	medication EPO	7.8	0.0
	medication oral iron	5.3	0.6
	medication other	5.3	0.6
	medication parenteral iron	4.1	0.3
	medication vitamin D	5.3	0.6
	weekly alcohol intake	43.2	4.3

Table 6: Summary statistics, over all follow-up years, for continuous variables before and after imputation

		Before				After			
item	units	min	max	median	IQR	min	max	median	IQR
general									
anti-HT *		0.0	8.0	2.0	2.0	0.0	8.0	2.0	2.0
BMI	kg/m^2	13.3	65.3	27.9	7.6	13.3	65.3	27.8	7.6
DBP	mmHg	40.5	141.5	72.5	15.0	40.0	142.0	72.0	15.0
PP	mmHg	17.0	188.0	63.0	26.5	17.0	188.0	63.0	27.0
SBP	mmHg	76.0	255.0	137.0	28.0	76.0	281.0	137.0	28.0
biochemical									
CC	mmol/L	1.0	3.3	2.3	0.2	1.0	3.3	2.3	0.2
CHO	mmol/L	2.1	16.0	4.3	1.4	1.9	16.0	4.3	1.4
CO2	mmol/L	6.0	44.5	22.8	4.7	6.0	44.5	22.8	4.5
CRP	mg/L	0.1	471.5	3.4	6.4	0.0	471.5	3.3	6.1
Hb	g/L	10.9	204.0	123.0	22.0	11.0	220.0	123.0	22.2
HbA1c	mmol/mol	24.6	159.0	48.6	22.8	24.6	192.2	44.3	19.4
PO	mmol/L	0.2	4.2	1.1	0.3	0.2	4.2	1.1	0.3
PTH	pmol/L	0.2	250.4	8.1	9.7	0.1	250.4	7.6	9.1
Pu †	g/24hr	0.0	18.5	0.3	0.8	0.0	18.5	0.3	0.8

* number of antihypertensives

† Due to rounding minimum Pu displays as 0.0 whereas before and after imputation it is actually 0.02.

2.8 Baseline variables

There are a number of reasons that a variable may only be present at baseline e.g. it never changes over time or was only recorded at the first appointment. However in some instances due to the sparseness of data we reduced a variable to a baseline value using the first recorded instance of the variable in the patient's data. For example, if the variable was not recorded at baseline but was instead recorded at the first follow-up appointment we used this value as if it were recorded at baseline. Variables reduced to baseline variables were: occupation, smoking status and weekly alcohol intake.

3 Linear mixed effects model

We have longitudinal data where each experimental unit (patient) consists of temporally correlated measurements over consecutive follow-up years. Classic multivariate models are not appropriate for analysing this grouped and correlated data. Standard extensions, for longitudinal data, to classical statistical procedures which estimate the parameters in regression models include the Generalised Estimating Equations (GEEs) (e.g. see (38,64)) and mixed effects models. A GEE is used to estimate the parameters of a generalised linear model. Specifically it aims to estimate the average response over the population rather than the regression parameters, the latter enables prediction of the effect of changing one or more explanatory variables on a given unit. GEEs are a widely used alternative to the likelihood-based mixed effects model which have the disadvantage of being more sensitive to the specification of the variance structure. However in our context we rejected the GEE approach because it is not robust to missing data due to patients missing follow-up appointments and/or spend differing lengths of time in the study. Our data contains both of these characteristics in abundance so we turn our attention to mixed effects models as they are able to accommodate this variability. In general mixed effects models are a commonly used class of statistical models that are applicable to a wide range of data structures which include correlated and/or clustered observations, repeated measurements and longitudinal measurements. It is not uncommon for longitudinal data to be modelled with mixed effects models consequently there exists an extensive literature; for example see texts (36–39,65).

Mixed effects models consist of both *fixed effects* and *random effects*, they describe the relationships between an outcome variable and explanatory variables. Fixed effects are associated with the whole population. There can be one or more layers of random effects when the data are grouped according to one or more classification levels. In this thesis we associate the random effects with individual experimental units drawn at random from a population. This model allows for clear identification of both population and individual patient characteristics. From this point onwards we consider only linear mixed effects (LME) models where the outcome variable is described by a linear function of the parameters.

Given the dataset described in Section 2 the data are sub-divided into disease categories and grouped at patient level. The LME model outcome variable is log(eGFR) and all the remaining variables are potential explanatory variables. In this thesis the combination of fixed effects plus random effects is interpreted as representing the unobserved GFR, therefore the LME model will express eGFR as a noisy version of GFR.

Event data which describe patients leaving the study (dropout, RRT or death) are not explicitly included in the model as we assume these events are missing at random; we did not test this assumption. It was beyond the scope of this thesis to explore models, e.g. survival models, which include this time to event data. For reviews relating to event data in the context CKD and mixed

effects models see for example (41,45).

We consider the following LME model for longitudinal trajectories given $i = 1, \ldots, M$ patients and $j = 1, \ldots, n_i$ observations per patient

$$Y_{ij} = \mu_i(t_{ij}) + U_i(t_{ij}) + \epsilon_{ij}. \tag{2}$$

The outcome for patient i at time $t_{ij} \geqslant 0$ is denoted Y_{ij}. The time since baseline measurement is t_{ij}, both n_i and t_{ij} vary among patients. This allows us to include patients with intermittent missing data and/or dropout, and also account for the actual individual measurement times. The expected value of the outcome is a multiple linear regression of the form $\mu_i(t_{ij}) = \boldsymbol{X}_i(t_{ij})\boldsymbol{\beta}$. Term $\mu_i(t_{ij})$ captures the fixed effects with a set of known explanatory variables $\boldsymbol{X}_i$ ($n_i \times p$ regressor matrix) and corresponding set of unknown fixed effects regression parameters $\boldsymbol{\beta}$ (p-dimensional vector) which are to be estimated. We assume any measurement errors in the explanatory variables are very much less than ϵ_{ij}.

The variability between patients which cannot be explained by the fixed effects is captured by the random effects described by a second linear regression $U_i(t_{ij}) = \boldsymbol{X}_i^*(t_{ij})\boldsymbol{b}_i$ with a known regressor matrix $\boldsymbol{X}_i^*$ (size $n_i \times q$) and corresponding vector of unknown random variables $\boldsymbol{b}_i$ (size q-dimensional vector) which are to be estimated. The distribution of $\boldsymbol{b}_i$ are assumed to be mutually independent multivariate normal random variables with mean zero, that is $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is a symmetric positive definite (non-degenerate) matrix hence is invertible. In particular we choose an intercept-and-slope model, the so-called Laird and Ware model (39), as such $\boldsymbol{X}_i^*(t_{ij}) = (\boldsymbol{1}_{n_i}, \boldsymbol{t}_i)$ where n_i-dimensional vector $\boldsymbol{t}_i$ has elements t_{ij}. The first term does not depend on time so represents the time-constant differences between patients and the second term represents the time dependent differences (variations in linear slope) between patients.

Random variables ϵ_{ij} are mutually independent with $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, given outcome Y_{ij} they account for the fact that eGFR is a noisy estimate of GRF. We refer to $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ij}, \ldots, \epsilon_{in_i})^T$ as within-group errors therefore without placing further constraints on Equation 2 it follows that $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_{n_i})$ where $\boldsymbol{I}$ denotes identity matrix. The errors are assumed to be independent for different groups (patients); independent of repeated measurements within the same group i; independent of random effects $\boldsymbol{b}_i$; and homoscedastic, that is having constant variance for both different groups and repeated measurements within the same group.

Given repeated measurements on patient i it may be necessary to take into account the correlation and variance of within-group errors to explain the change over time of outcome Y_{ij} not explained by the aforementioned linear regressions. To this end let

$$\boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Lambda}_i) \tag{3}$$

with variance-covariance matrix $\boldsymbol{\Lambda}_i$. This matrix is symmetric positive definite and decomposed such that

$$\boldsymbol{\Lambda}_i = \boldsymbol{V}_i \boldsymbol{C}_i \boldsymbol{V}_i. \tag{4}$$

The variance matrix $\boldsymbol{V}_i$ is diagonal and the correlation matrix $\boldsymbol{C}_i$ has diagonal elements equal to one. This decomposition therefore allows the variance and correlation structures of the within-group errors to be modelled separately. It follows that

$$\text{var}(\epsilon_{ij}) = \sigma^2 [\boldsymbol{V}_i]^2_{jj} \tag{5}$$

and

$$\text{cor}(\epsilon_{ij}, \epsilon_{ij'}) = [\boldsymbol{C}_i]_{jj'} \tag{6}$$

with $j' = 1, \ldots, n_i$. Hence the correlation structure accounts for repeated measurements within group i. This formulation assumes $\boldsymbol{\epsilon}_i$ is independent for different groups i and independent of random effects $\boldsymbol{b}_i$. In our study we assume the variance structure is homoscedastic $\text{var}(\epsilon_{ij}) = \sigma^2$ as we found no evidence to the contrary, therefore in the following we will now focus on the correlation structure. The correlation between two within-group errors ϵ_{ij} and $\epsilon_{ij'}$ is assumed to depend on the magnitude of their temporal distance. In particular the correlation structure is assumed to be isotropic so it depends only on relative distances and not the temporal positions. This distance is described by the function $\delta = d(\boldsymbol{p}_{ij}, \boldsymbol{p}_{ij'})$ where $\boldsymbol{p}_{ij}$, $\boldsymbol{p}_{ij'}$ are position vectors for ϵ_{ij}, $\epsilon_{ij'}$ respectively. With reference to Equation 6 let the correlation structure be defined by

$$\text{cor}(\epsilon_{ij}, \epsilon_{ij'}) = h(\delta, \boldsymbol{\rho}) \tag{7}$$

where autocorrelation function $h(\cdot)$ takes values between -1 and 1 and $\boldsymbol{\rho}$ is a vector of correlation parameters. Note 1: if we assume no correlation structure then $h(\cdot)$ will be zero everywhere except on the diagonal. Note 2: $h(\cdot)$ is defined such that if the distance between the position vectors is zero then $h(0, \boldsymbol{\rho}) = 1$. Given repeated measurements on each patient i a natural choice of correlation structure would be a zero mean continuous-time stochastic process, such as a first order continuous-time autoregressive model (CAR1). This model is defined by $h(s, \rho) = \rho^s$ where $\rho \geqslant 0$ and the magnitude of the time difference $s \geqslant 0$ (e.g. $s = |t_{ij+1} - t_{ij}|$). It can be seen that the correlation function decreases in absolute value exponentially with decay constant $\tau = -1/\ln\rho$ since $h(s, \rho) = \text{e}^{s \ln \rho} = \text{e}^{-s/\tau}$; i.e. events close together are more correlated than distant events.

Alternatively given many patients have very few follow-up measurements (see Figures 54 to 62) a compound symmetry (CS) structure may be more suitable, as suggested by Pinheiro and Bates (66) (see Chapter 5) who state that CS may be useful if each group's timeseries is short. The CS model is defined as $0 \leqslant \rho \leqslant 1$ with $h(k, \rho) = \rho \; \forall j \neq j'$ otherwise $h(k, \rho) = 1$; integer time differences are denoted by $k = 1, 2, \ldots,$. In Section 5.4 we investigate whether there is sufficient evidence to include a correlation structure in our models.

To fit the model in Equation 2 when $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \; \boldsymbol{I}_{n_i})$ we need to estimate $\boldsymbol{\beta}$, $\boldsymbol{b_i}$, $\boldsymbol{\Psi}$ and σ. If we find enough evidence for within-group error correlations then $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$ hence additional parameters associated with $\boldsymbol{C}_i$ will need estimating. We fit these LME models within the maximum likelihood framework using R-package nlme (66,67). This approach uses the conditional modes of the random effects given the data. A full mathematical description is given in Chapter 2 of (66).

4 Inferences regarding changes in eGFR

Our primary interest is to determine, for each disease model, the degree to which the fixed effect explanatory variables explain the outcome at population level. As is usual we inspect each regression parameter value along with its corresponding statistical significance when reporting the results in Chapter 7. However these results are relative to log(eGFR), it is not possible to directly interpret them in terms of eGFR which is the unit that clinicians are typically familiar with. As a consequence results expressed in log(eGFR) are not fully accessible to the intended audience of this research; for example a clinician may be interested in the benefits in terms of eGFR of prescribing a given medication. In Section 4.1 we address this by introducing methodology to assess the average effect on eGFR of a small step change in a given explanatory variable; we make use of this when reporting our results in Chapter 7. It may also be of interest to interpret the model in terms of how quickly the model outcome is on average changing over time, therefore in Section 4.2 we introduce methodology for investigating the temporal trajectory of both log(eGFR) and eGFR.

With reference to Equation 2 we rewrite the fitted LME model in component form, with intercept-and-slope random effect, as follows

$$\hat{Y}_{ij} = \hat{\beta}_0 + \sum_{r=1}^{p-1} \hat{\beta}_r X_i^{(r)}(t_{ij}) + \hat{b}_{i0} + \hat{b}_{i1} t_{ij} \tag{8}$$

where the model parameters have been estimated by maximum likelihood. The intercept and slope random effects terms are defined respectively as $\hat{b}_{i0}$ and $\hat{b}_{i1}$. The outcome $\hat{Y}_{ij}$ represents $\log_e(e_0^{-1}\text{eGFR}(t))$. The constant $e_0 = 1$ has units identical to eGFR, this ensures the argument of the logarithm does not carry physical dimensions (units). The outcome in terms of eGFR(t) is

$$\hat{Y}^*_{ij} = e_0 e^{\hat{Y}_{ij}}. \tag{9}$$

4.1 Step changes in explanatory variables

4.1.1 Step changes on log(eGFR) scale

A standard interpretation of Equation 8 is that if we hold all terms constant except one, e.g. variable $X_i^{(r)}(t_{ij})$, then for every additional increase of one unit in $X_i^{(r)}(t_{ij})$ we expect the outcome to change by an average of $\hat{\beta}_r$. In other words given a change from $X_i^{(r)}(t_{ij})$ to $X_i^{(r)}(t_{ij}) + \theta_r$, we define $\Delta^r \hat{Y}_i = \hat{Y}^\theta_{ij} - \hat{Y}_{ij}$ where $\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_r X_i^{(r)}(t_{ij}) + \ldots$ and $\hat{Y}^\theta_{ij} = \hat{\beta}_0 + \hat{\beta}_r(X_i^{(r)}(t_{ij}) + \theta_r) + \ldots$; therefore for the r^{th} regression term $\Delta^r \hat{Y}_i = \theta_r \hat{\beta}_r$. The term $\Delta^r \hat{Y}_i$ describes the amount $\hat{Y}_{ij}$ shifts

when subjected to a change of size θ_r. Given θ_r is a constant over time then $\hat{Y}_{ij}$ and $\hat{Y}_{ij}^{\theta}$ have identical time derivatives therefore a step change of size θ_r affects only the value of log(eGFR) and not its rate of change. If θ_r is applied to the r^{th} explanatory variable across all i patients, it follows that on average log(eGFR) changes by

$$\Delta^r \hat{Y} = \theta_r \hat{\beta}_r. \tag{10}$$

Note that $\Delta^r \hat{Y}_i$ and $\Delta^r \hat{Y}$ are dimensionless.

We could set $\theta_r = 1$ for all explanatory variables but given there are orders of magnitude differences between our variables this could be very misleading when assessing the degree to which each explanatory variable contributes to changes in either log(eGFR) or eGFR. In practice we suggest assigning a value to θ_r which is commensurate with a typical change in the explanatory variable of interest. One possibility, for the r^{th} explanatory variable from all patients, would be to set θ_r equal to the mean of the differences in the absolute value between successive follow-up appointments; i.e. find the mean of $|X_i^{(r)}(t_{ij}) - X_i^{(r)}(t_{i\,j+1})|$ over all i and j. However in this thesis we use the standard statistical approach of setting θ_r equal to one standard deviation of the distribution of observations from the r^{th} explanatory variable; i.e. for a given r, one standard deviation of the distribution of $X_i^{(r)}(t_{ij})$ over all i and j. The exception is categorical variables which always have $\theta_r = 1$. Furthermore if (non-categorical) explanatory variables are standardised then for each such variable $\theta_r = 1$. Note that standardisation is the process of putting the variables on the same scale, in this thesis standardisation is performed for each variable by subtracting the mean and dividing by the standard deviation.

4.1.2 Step changes on eGFR scale

We now extend the ideas in Section 4.1.3 to estimating changes in eGFR as opposed to log(eGFR). Specifically we want to determine how eGFR varies given a change of size θ_r in an explanatory variable. We considered three approaches for estimating this change:

- *Proportional change*, this is obtained by directly transforming $\Delta^r \hat{Y}_i$ (see Equation 10) to the eGFR scale as follows:

$$\begin{aligned} e^{\Delta^r \hat{Y}_i} &= e^{\hat{\beta}_r \theta_r} \\ &= e^{\hat{Y}_{ij}^{\theta} - \hat{Y}_{ij}} \\ &= \hat{Y}_{ij}^{*\theta} / \hat{Y}_{ij}^{*} \end{aligned} \tag{11}$$

  where $\hat{Y}_{ij}^{*\theta} = e_0 e^{\hat{Y}_{ij}^{\theta}}$ and $\hat{Y}_{ij}^{*} = e_0 e^{\hat{Y}_{ij}}$. This is a ratio in eGFR, i.e. $\hat{Y}_{ij}^{*\theta} / \hat{Y}_{ij}^{*}$, that is the proportional change in eGFR induced by a change of size θ_r. We will not use this approach

when reporting results as we seek a quantity which represents the difference (not a ratio) in eGFR induced by a change in θ_r. Two such approaches are given in the following two bullet points.

- *Absolute difference*, this is obtained by first considering the expression $\Delta^r \hat{Y}^*_{ij} = \hat{Y}^{*\theta}_{ij} - \hat{Y}^*_{ij}$. Writing this out in full we obtain $\Delta^r \hat{Y}^*_{ij} = e_0 \exp(\hat{\beta}_0 + \hat{\beta}_r(X^{(r)}_i(t_{ij}) + \theta_r) + \dots) - e_0 \exp(\hat{\beta}_0 + \hat{\beta}_r X^{(r)}_i(t_{ij}) + \dots)$, from which it follows that

$$\Delta^r \hat{Y}^*_{ij} = \hat{Y}^*_{ij}(e^{\hat{\beta}_r \theta_r} - 1). \tag{12}$$

  As such we can assess the effect of θ_r on $\Delta^r \hat{Y}^*_{ij}$. The absolute difference in eGFR at population level could be defined as

$$\mathbb{E}(\Delta^r \hat{Y}^*) = \mathbb{E}(\hat{Y}^*_{ij})(e^{\hat{\beta}_r \theta_r} - 1) \tag{13}$$

  where $\mathbb{E}(\hat{Y}^*_{ij})$ is the expected value of $\hat{Y}^*_{ij}$ over the population and all time. For our dataset $\mathbb{E}(\hat{Y}^*_{ij}) = 31.5$ mL/min/1.73m^2. However a shortcoming of this approach is that the value of $\mathbb{E}(\hat{Y}^*_{ij})$ is dataset specific and $\hat{Y}^*_{ij}$ is highly variable across the population. We therefore do not report results using this approach.
- *Relative change* in eGFR, given Equation 12, is defined as

$$\Delta^r \hat{Y}^* = \Delta^r \hat{Y}^*_{ij} / \hat{Y}^*_{ij} = e^{\hat{\beta}_r \theta_r} - 1 \tag{14}$$

  This approach is not subject to the aforementioned shortcomings therefore we use it when reporting results in section 7.

Note that $\hat{Y}^*_{ij}$, $\Delta^r \hat{Y}^*_{ij}$ and $\Delta^r \hat{Y}^*$ have the same physical dimensions as eGFR.

4.1.3 Summary of Step changes approaches

Given clinicians typically work on the eGFR scale, and not on the log scale, we report our results relating to step changes in θ_r using the relative change approach given in Equation 14. As described in section 4.1.3 we use θ_r equal to one standard deviation of the r^{th} explanatory variable distribution, i.e. the distribution of $X^{(r)}_i(t_{ij})$ over all i and j. It follows that if this distribution is standardised then the step size will equal one since the standard deviation is one.

4.2 Rates of change over time

The LME model given in Equation 2 has an error term $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, as already discussed. This term may have within-group correlations described by a stochastic process such as the aforementioned CAR1 model. The time derivative of Equation 2 would necessarily need to account for the stochasticity of the error term. However it is beyond the scope of this current work to

consider fitting such models. Here we circumvent this issue by focusing on the time derivative of the fitted model i.e. the derivative of Equation 8.

The trajectory of explanatory variable $X_i^{(r)}(t_{ij})$ through time may be constant, continuous or piecewise continuous:

- Each baseline explanatory variable, e.g. ethnicity, is constant over all time hence its time derivative is zero.
- Each explanatory variable which changes smoothly over time, e.g. biochemical markers, are continuous functions of time. Although we only have observations at fixed points in time we may interpolate, e.g. with a spline, between observations; hence the spline's derivative represents the variable's time derivative.
- Each categorical variable which varies over time is a piecewise continuous function in time. The derivative of such a variable exists everywhere except at time points where it changes level; at these points there exists a discontinuity. Outside of the discontinuities the variable is constant with respect to time hence its derivative is zero.
- In this section we consider interaction terms of the form $t_{ij}X_i^{(r)}(t_{ij})$ to be a special case of $X_i^{(r)}(t_{ij})$. An interaction term between time and a categorical variable is piecewise continuous function of time whose derivative exists everywhere except where the categorical variable changes levels; outside of the discontinuities the time derivative[1] of $t_{ij}X_i^{(r)}(t_{ij})$ equals $X_i^{(r)}(t_{ij})$.

4.2.1 Time derivative on log(eGFR) scale

With reference to Equation 8 we seek the time derivative of $\log_e(e_0^{-1}\text{eGFR}(t))$ i.e.

$$\dot{\hat{Y}}_{ij} = \sum_{r=1}^{p-1} \hat{\beta}_r \dot{X}_i^{(r)}(t_{ij}) + \hat{b}_{i1} \tag{15}$$

where dot denotes the first order time derivative e.g. $\dot{X} = dX/dt$. We assume $X_i^{(r)}(t_{ij})$ can be represented by a continuous function which is differentiable. Time independent and categorical variables essentially have time derivatives of zero. The regression parameters are not estimated from Equation 15. They are estimated in the usual way, as described in Chapters 3 and 5, including those whose corresponding explanatory variable has a time derivative of zero in Equation 15.

The additive nature of Equation 15 allows us to focus on the r^{th} regression term of patient i; its contribution to the outcome $\dot{\hat{Y}}_{ij}$ is denoted

$$\dot{\hat{Y}}_{ij}^{(r)} = \hat{\beta}_r \dot{X}_i^{(r)}(t_{ij}). \tag{16}$$

[1]Note: $\frac{d}{dt}(t.X(t)) = t.\dot{X}(t) + \dot{t}.X(t) = t.0 + 1.X(t) = X(t)$.

We do not compute $\dot{X}_i^{(r)}(t_{ij})$ using a statistical model, for example an intercept-and-slope linear model, as estimation of the LME model parameters in Equation 2 assumes explanatory variable observations exhibit negligible noise (e.g. measurement error) compared with the error terms ϵ_{ij}. Here we calculate $\dot{X}_i^{(r)}(t_{ij})$ by performing a cubic spline interpolation around the explanatory variable's data points, and then compute the spline's time derivative which we denote $\dot{\mathbb{S}}_i^{(r)}(X_t)$ where $X_t \equiv X_i^{(r)}(t_{ij})$. Hence Equation 16 is approximated by

$$\dot{\hat{Y}}_i^{(r)}(t) = \hat{\beta}_r \dot{\mathbb{S}}_i^{(r)}(X_t). \tag{17}$$

The average trajectory is $\hat{\xi}_i^{(r)} = \frac{1}{T}\int_T \dot{\mathbb{S}}_i^{(r)}(X_t)dt$ hence Equation 17 is then

$$\dot{\hat{Y}}_i^{(r)} = \hat{\beta}_r \hat{\xi}_i^{(r)}. \tag{18}$$

At population level, the average rate of change over time of the r^{th} explanatory variable is estimated by taking its expected value over all patients

$$\mathbb{E}(\dot{\hat{Y}}_i^{(r)}) = \hat{\beta}_r \mathbb{E}(\hat{\xi}_i^{(r)}). \tag{19}$$

Moreover the distribution of all $\dot{\hat{Y}}_i^{(r)}$ for the r^{th} explanatory variable can be used to estimate confidence intervals.

Similarly we estimate the average trajectory over time of the outcome variable, log(eGFR), as follows. Given Equation 15 for patient i, we perform spline interpolation on all regression terms, then sum over all terms and finally calculate the i^{th} average trajectory by integrating over time. The population's overall trajectory is then the expected value of all the i^{th} average trajectories, which we denote $\mathbb{E}(\dot{\hat{Y}}_i)$.

Note that $\dot{Y}_{ij}^{(r)}$, $\mathbb{E}(\dot{\hat{Y}}_i^{(r)})$ and $\mathbb{E}(\dot{\hat{Y}}_i)$ have dimensions of one over time. In our study the unit of time is a year.

4.2.2 Time derivative on eGFR scale

It follows from Equations 8, 9 and 15 that the time derivative in terms of eGFR(t) for patient i is[2]

$$\dot{\hat{Y}}_{ij}^{*} = \hat{Y}_{ij}^{*}\,(\sum_{r=1}^{p-1} \hat{\beta}_r \dot{X}_i^{(r)}(t_{ij}) + \hat{b}_{i1}). \tag{20}$$

[2]Note: $\frac{d}{dt}\log_e(f(t)) = \dot{f}(t)/f(t)$.

The influence of a single term, e.g. $\hat{\beta}_r \dot{X}_i^{(r)}(t_{ij})$, on the outcome for patient i is given by

$$\dot{\hat{Y}}_{ij}^{*(r)} = \hat{Y}_{ij}^{*} \; \hat{\beta}_r \dot{X}_i^{(r)}(t_{ij}). \tag{21}$$

For patient i, as above performing spline interpolation, leads to $\dot{\hat{Y}}_{i}^{*(r)}(t) = \mathbb{S}_i(\hat{Y}^*) \; \hat{\beta}_r \dot{\mathbb{S}}_i^{(r)}(X_t)$; given patient i then $\mathbb{S}_i(\hat{Y}^*)$ denotes the spline interpolation of the outcome's fitted values. Given the average trajectory $\hat{\xi}_i^{*(r)} = \frac{1}{T}\int_T \mathbb{S}_i(\hat{Y}^*) \; \dot{\mathbb{S}}_i^{(r)}(X_t)dt$ then Equation 21 is estimated with

$$\dot{\hat{Y}}_{i}^{*(r)} = \hat{\beta}_r \hat{\xi}_i^{*(r)}. \tag{22}$$

The analogue at population level is given by the expected value of $\hat{\xi}_i^{*(r)}$ over all i

$$\mathbb{E}(\dot{\hat{Y}}_{i}^{*(r)}) = \hat{\beta}_r \mathbb{E}(\hat{\xi}_i^{*(r)}) \tag{23}$$

and distribution of all $\dot{\hat{Y}}_{i}^{*(r)}$ will be used to estimate confidence intervals. In the results section 7.4 we report rates using Equation 23 and corresponding confidence intervals based on a bootstrap method which does not assume a normal distribution.

Given Equation 20 the expected average trajectory of the outcome, eGFR, for the population, denoted $\mathbb{E}(\dot{\hat{Y}}_{i}^{*})$, is estimated as previously described (see paragraph after Equation 19) i.e. population's overall trajectory is then the expected value of all the i^{th} average trajectories. This quantity is also reported in the results section 7.4.

Note that $\dot{\hat{Y}}_{ij}^{*(r)}$, $\mathbb{E}(\dot{\hat{Y}}_{i}^{*(r)})$ and $\mathbb{E}(\dot{\hat{Y}}_{i}^{*})$ have units of eGFR per unit time.

4.3 Interpreting sign of regression parameters in terms of temporal progression

Here we rewrite Equation 8 with an explicit fixed effect explanatory variable for time, that is

$$\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 t_{ij} + \sum_{r=2}^{p-1} \hat{\beta}_r X_i^{(r)}(t_{ij}) + \hat{b}_{i0} + \hat{b}_{i1} t_{ij}. \tag{24}$$

In our data all continuous explanatory variables always have positive values. We focus on the first three terms of Equation 24 and rewrite it in terms of eGFR(t) as follows $\hat{Y}_{ij}^{*}(t) = e_0 \exp(\hat{\beta}_0 + \hat{\beta}_1 t_{ij} + \hat{\beta}_2 X_i^{(2)}(t_{ij}) + \dots)$. The prefactor $e_0 \exp(\hat{\beta}_0)$ determines the intercept at $t = 0$. The middle term $\exp(\hat{\beta}_1 t_{ij})$ with $\hat{\beta}_1 < 0$ gives an exponential rate of decay of eGFR(t) over time, hence larger values of $|\hat{\beta}_1|$ result in faster decay rates: consequently kidney function deteriorates

more rapidly. If $\hat{\beta}_1 > 0$ this would indicate an improvement in kidney function. The last term $\exp(\hat{\beta}_2 X_i^{(2)}(t_{ij}))$ will indicate decreasing eGFR(t) over time when $\hat{\beta}_2 < 0$ and $x_i^{(2)}(t) > 0$ is monotonically increasing over all time. Likewise kidney function will be worsening if $\hat{\beta}_2 > 0$ and $x_i^{(2)}(t) > 0$ is monotonically decreasing. Consequently the sign of the regression parameter and the explanatory variable's trajectory over time determine whether the regression term contributes towards an improvement or deterioration in kidney function.

4.4 Interpretion of fixed effects temporal interaction terms

With respect to the log(eGFR) model and its time derivative we consider the interpretation of the fixed effects interaction terms. In this thesis all interactions are with follow-up time.

4.4.1 Regression model for log(eGFR)

We use interaction terms between a given explanatory variable and follow-up time, which we denote by *explanatoryVariable : followupTime*; in mathematical notation this is may be written $x(t)\,t$. For every interaction term we also include the corresponding explanatory variables as separate terms for example $\beta_1 t + \beta_2 x(t) + \beta_3 x(t)\,t$, hence rearranging gives $\beta_1 t + (\beta_2 + \beta_3 t)x(t)$. The factor $(\beta_2 + \beta_3 t)$ describes the time-independent (β_2) and time-dependent $(\beta_3 t)$ effects on $x(t)$.

4.4.2 Regression model for rate of change in log(eGFR) over time

A regression model for the rate of change over time of outcome eGFR will be computed by taking the time derivative of terms such as $\beta_1 t + \beta_2 x(t) + \beta_3 x(t)\,t$, the time derivative of this expression is $\beta_1 + (\beta_2 + \beta_3\,t)\dot{x}(t) + \beta_3 x(t)$. Similarly to Section 4.4.1 the factor $(\beta_2 + \beta_3\,t)$ describes the time-independent (β_2) and time-dependent $(\beta_3 t)$ effects on $\dot{x}(t)$ however there is an additional time-dependent effect through the $\beta_3 x(t)$ term. If $x(t)$ is a categorical variable then $\dot{x}(t) = 0$ everywhere except at any discontinuities where it is undefined; therefore $\beta_1 + (\beta_2 + \beta_3\,t)\dot{x}(t) + \beta_3 x(t)$ reduces to $\beta_1 + \beta_3 x(t)$ hence in terms of this rates of change model β_2 has no effect. Another way of looking at this is when $y(t) = \beta_1 t + \beta_2 x(t) + \beta_3 x(t)\,t + \ldots$ is differentiated with respect to time, i.e. $\dot{y}(t) = \beta_1 + \beta_2 \dot{x}(t) + \beta_3 d(x(t)\,t)/dt + \ldots$, the parameters quantify the rate of change of log(eGFR) per unit time (year). Although we do not fit the differentiated model this interpretation stands.

4.5 Standardised model

From this point onwards, unless otherwise stated, all regression models will use standardised continuous explanatory variables. The rationale being that this will allow us to assess the relative

importance of the fixed effects regression parameters once the model is fitted. To standardise each variable we subtract its mean and divide by its standard deviation. Standardisation is a widely used technique when comparing model parameters but is open to criticism, for example the meaning of one standard deviation may be open to debate especially for small sample sizes or non-normal distributions. In this thesis we consider standardisation to be a pragmatic method of rescaling the continuous explanatory variables to the same scale. The standardised variables are dimensionless (no units of measure).

To aid interpretation follow-up time and baseline age are not standardised hence retain their units of time i.e. years. Given follow-up time is not standardised the model can still be interpreted in relation to disease progression per year. Furthermore the outcome variable log(eGFR) is not standardised.

With reference to Section 4.1 when the variables are standardised a unit step change in the standardised explanatory variable results in a one standard deviation change in the (unstandardised) variable of interest. It follows that, with the standardised quantities denoted by dash, then $\beta_r\sigma_r(X_r/\sigma_r + \theta/\sigma_r) = \beta_r^{'}(X_r^{'} + \theta_r^{'})$ therefore $\theta_r^{'} = \theta_r/\sigma_r$; i.e. if $\theta_r^{'} = 1$ then $\theta_r = \sigma_r$.

5 Model selection

First we checked if there existed any significant dependence between the risk factors. To identify the factors in our models which best describe the progression of kidney disease for each disease we used a bi-directional selection procedure (based on the Akaike information criterion) on multiple bootstrap samples; this allows us to gauge parameter uncertainty and helps to guard against overfitting to the SKS data.

5.1 Dependence among model variables

In our regression models we need to avoid multicollinearity, that is the phenomenon by which one variable can be linearly predicted from other variable(s) with a substantial degree of accuracy. Multicollinearity occurs when two or more covariates are highly correlated which leads to unreliable and unstable estimates of regression parameters.

To assess the strength of correlation between all pairs of covariates we computed the correlation matrix; results are tabulated in Tables 31 to 36 of Appendix A.4.1. We did not find any unexpectedly strong correlations. As expected covariates which were computed from, or strongly related to, other covariates had strong correlations in particular: log(eGFR) and Cr; PP and SBP; past cancer and no cancer.

To detect multicollinearity among covariates we used the variance inflation factor (VIF) which is one of the most widely used methods (68). VIF is calculated for each covariate by performing a linear regression of that covariate on all the other covariates, and then obtaining the coefficient of determination R^2 from that regression. VIF for a given covariate is defined as $1/(1-R^2)$ and has a range from 1 upwards where 1 indicates the covariate is completely uncorrelated with all other covariates. Hence VIF estimates how much the variance of a regression coefficient is inflated due to its covariate's association with all the other covariates; for example if the VIF is 1.9 then the variance of the given regression coefficient is 90% larger than would be expected if its associated covariate was completely uncorrelated with all the others. To compute VIF values for all potential covariates we employed an algorithm which uses a stepwise procedure, in particular we use function vifstep from R-package usdm (69). First, the algorithm calculated the VIF for every variable, then it excluded the variable with the highest VIF provided its VIF exceeded a predefined threshold, this procedure was repeated until there were no remaining variables with a VIF greater than the threshold. It is generally agreed that a VIF greater than 10 indicates too much multicollinearity (e.g. Section 9.4 in (68)) but some authors consider there is too much if VIF is higher than 5 and others if higher than 2.5; for example see discussion by (70). For our data with a threshold set at 5 this method excluded SBP, reducing the threshold to 2.5 then resulted in the exclusion of the *number of antihypertensives* patients were taking. In addition the indicator variables derived from the categorical variable for disease also showed some high VIF

values. Appendix A.4.2 lists the excluded variables and tabulates VIF values for variables whose VIF values are less than the aforementioned thresholds; see Tables 37 and 38.

In conclusion, having assessed the strength of dependence between variables we decided to exclude SBP as the clinicians advised their preferred blood pressure measure in the context of this research was PP. Our VIF analysis indicates that with a threshold of 2.5 we should consider excluding the *number of antihypertensives* however we opt to use this variable in our models as the SKS clinicians advise us of its importance. In this sense we are effectively using a VIF with a threshold of 5. Some of the indicator variables derived from the categorical variable for disease had relatively high VIF values however there was an indicator variable for every disease category so some degree of correlation or anti-correlation is to be expected therefore multicollinearity in this context it is not a cause for concern. Given log(eGFR) is the outcome variable in our models we will not use Cr as a covariate, the strong correlation between the two arises because Cr is used to compute eGFR (Equation 1); if Cr was included it could obscure the effects from other variables which are our primary interest. We note sex and ethnicity were not strongly correlated with log(eGFR). This is probably because sex is a small effect in the eGFR formula (Equation 1) and although ethnicity is a slightly larger effect the vast majority of the SKS cohort were classified as 'white' thereby obscuring any strong association.

5.2 Stepwise regression with bidirectional selection and bootstrapping

With the exception of considering dependencies among covariates in the previous section all the covariate selection has up to this point in the thesis been based on the guidance and expertise of the renal clinicians at Salford Royal NHS Foundation Trust who designed the SKS study. This expert knowledge is invaluable for assisting with model selection, but creating statistical models with a large number of covariates, as we have here, could potentially lead to overfitting. An overfitted model would describe some of the residual variation (noise) as if this variation represented part of the underlying model structure or physical process. Hence such models exaggerate minor fluctuations in the data. Usually there is a trade-off between goodness-of-fit and parsimony since models with many parameters tend to have a better model fit to the data but will perform poorly when predicting from other datasets.

Our objective is to create parsimonious models; the simplest models with the least number of covariates but with greatest explanatory power. There are various methods to estimate the balance between parsimony and goodness-of-fit, popular methods include:

- Akaike's Information Criterion, AIC - introduced by Akaike 1973 (71–73) - given the number of estimated parameters k and the maximum value $\hat{L}$ of the likelihood function of a candidate model then $AIC = 2k - 2\ln\hat{L}$. Hence AIC rewards goodness-of-fit as determined by the likelihood function but includes a penalty which increases with k that suppresses

overfitting. The best model from a set of candidate models is the one with the lowest AIC. Note that AIC does not describe model quality so given a set of poor models the AIC will select the best one from the poor-quality set.

- Bayesian Information Criterion, BIC - introduced by Schwarz 1978 (74) - uses a penalty term, similar to AIC, for the number of parameters in the model but the penalty term is larger hence BIC will often favour fewer parameters; $BIC = k\ln(n) - 2\ln\hat{L}$ where n is the number of data points.

Other popular methods include 'minimum description length' and 'Bayes factors'; for a description and comparison of these methods, see (75).

In this thesis we use AIC. First, AIC is considered asymptotically optimal for selecting the regression model (with the least mean squared error) from the set of candidates under the assumption that this set does not contain the 'true model' (i.e the process that generated the data). In contrast under this assumption BIC is not asymptotically optimal; see for example the comparison of AIC and BIC given by (76) in relation to regression models. Secondly, the risk of selecting a bad model is minimised with AIC compared with BIC which carries a significant risk of selecting a poor model from the candidate models; e.g. see simulation study by (77). Lastly, (77) suggests AIC is preferred when the 'true model' is complex relative to all candidate models, that is when all the candidates substantially oversimplify the underlying physical processes; this is most likely the case with our dataset as it is very doubtful we have all the required covariates to completely model the physical processes driving changes in renal function. It is also improbable that the complexities of renal function are fully described by the simple structure of our linear regression models.

To assist with model selection we used stepwise regression which is a method of fitting regression models in which the choice of covariates is carried out by a systematic procedure. In each step of the algorithm a covariate is considered for addition to, or subtraction from, the set of covariates based on AIC. We use, from the R-package MASS (78), the function stepAIC which is briefly described in (78) on page 175. This function implements a bidirectional selection procedure. To the author's knowledge neither (78) or the MASS documentation describe the algorithm so its steps are outlined here:

1. it computes AIC for the regression model with all covariates;
2. it removes each covariate one at a time (backward selection) from the regression model and calculates the AIC for each model then selects the one with lowest AIC;
3. it again removes covariates one at a time (backward selection) but also in turn adds covariates in one at a time which were previously removed (forward selection), then the regression model with the lowest AIC is selected;
4. the combination of backward-forward selection in step 3. is repeated until the model with the lowest AIC is found.

An exhaustive search where regression models are computed for every possible combination of covariates will find the global minima in AIC (or whichever statistic is used) but such a search is computationally impractical for the number of covariates in our dataset. The aforementioned bidirectional selection procedure, although typically more robust than applying only a forward or a backward selection procedure, still presents the risk of unknowingly selecting a model with a local minima in AIC rather than the desired model with the global AIC minima.

To gauge the level of model selection uncertainty we employed a bootstrap method; the principles of which were first published by Efron 1979 (79) and are now widely used e.g. see texts (80,81). This method, which is distribution-independent, is a resampling technique which estimates statistics on an unobserved population by sampling the observed dataset with replacement. In particular the observed dataset is randomly resampled with replacement, the bootstrap distribution is generated by repeating this resampling procedure a number of times. Provided the observed dataset is a representative sample from the true population the bootstrap method works by treating the true distribution as being analogous to the bootstrap distribution. It is therefore possible to assess the properties of the unobserved distribution of the population.

The bootstrapping technique typically assumes all observations are from an independent and identically distributed population. However this assumption is violated by longitudinal data. There are multiple observations per patient (cluster), and the data are independent between patients but temporally correlated within each patient's records. We respect this data structure by using the so-called m-out-of-n bootstrap where there are a total of n records grouped into m clusters; for example (81) page 140 and (82) discuss this type of bootstrap. In terms of our data the patients, i.e. m clusters, are randomly resampled with replacement while the observations for each patient remained unchanged so as to preserve temporal correlations. It follows that each bootstrap sample has the same number of patients (clusters) as the original data although some patients would almost surely occur more than once.

In summary, for a given dataset the final model will be obtained by using the bootstrap to estimate selection stability for each explanatory variable under bidirectional stepwise regression. The exact procedure is summarised below in Section 5.4.

5.3 Training and validation data

To help detect the presence of any under- or over-fitting in the aforementioned model selection procedures, described in Section 5.2, were performed on a subset of data, *training data*, and the resulting model was then validated using the remaining data, *validation data*. Commonly, training data consists of 75-80% of the entire dataset and the remaining 25-20% forms the validation data. In our case, for a given dataset, we obtained the training data by randomly selecting the desired number of patients (without replacement), therefore the remaining patient data formed

the validation data. The idea is that if the model selected fits similarly to both the training and validation data then we surmise that the model adequately describes the data without under- or over-fitting.

Note that the use of training and validation data is no more than a weak test of overfitting. For a discussion on its limitations see Section 8.3.

5.4 Summary of model selection procedure

We create a separate LME model for each primary kidney disease group (diabetic nephropathy, glomerulonephritis, hypertensive kidney disease, obstruction, other, polycystic kidney disease, pyelonephritis, renovascular disease, unknown) with the exception of obstruction which is excluded because of too little data. Additionally we make an overall model, called 'single model all diseases', which uses the entire dataset including patients with obstruction. We select our final models for each disease category as followings:

Step 1. Using the full dataset, strong correlations between covariates were eliminated by completely discarding several covariates; details given above in Section 5.1.

Step 2. Given Equation 2, for each disease we initially use a parsimonious LME model with random effect $\boldsymbol{X}_i^* = \mathbf{1}_{n_i}$ and $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\ \boldsymbol{I}_{n_i})$. Each model is fitted by maximising the log-likelihood so that we can compare models using AIC. Fixed effects for each disease model are selected as follows:

1. Wherever the dataset was large enough we apportioned 80% of patients (randomly selected) to the training data and the remaining 20% to the validation data. With this ratio the PKD and pyelonephritis disease models contained too few patients in the validation data so we apportioned 75% of patients to the training data and the remaining 25% to the validation data.
2. We generated 100 bootstrap samples from the training data.
3. The bidirectional model selection procedure was applied to each bootstrap sample. Given each bootstrap sample a regression model with the lowest AIC was estimated and its fixed effect regression parameters recorded.
4. We assessed regression model stability across all bootstrap samples by computing the proportion of samples in which each explanatory variable was included in the regression. The final model for each disease category was selected using explanatory variables which occurred in more than 50% of bootstrap samples.
5. The final models were fitted using the validation data to check the robustness of the model fit to the data.

Step 3. Given our interest is in the progression of disease over time, we augmented the fixed

effects with interaction terms between each time varying explanatory variable and time since follow-up. This allows us to estimate the effects on the slope of log(eGFR) over time. The rationale for not including interaction terms during *Step 2.* was to limit the size of the parameter space so as to reduce the chance of overfitting and/or selecting models in a local, rather than a global, minima. From this point onwards all models include such interactions which we denote as *expanatoryVariable:followupTime*. Note that we do not consider all possible interactions between all explanatory variables as again the parameter space would become too large potentially leading to sub-optimal models.

Step 4. The fixed effects selected in *Step 2. and 3.* were for a model with random effect design matrix $\boldsymbol{X}_i^* = \mathbf{1}_{n_i}$ and $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\ \boldsymbol{I}_{n_i})$; below we refer to this as 'Model A'. Using these fixed effects we investigated the model fit by undertaking rudimentary exploratory analysis using log-likelihood estimates. As is customary in longitudinal analysis we considered models with different random effects and correlation structures including compound symmetry (CS). The CS results are not presented here as they did not significantly improve the model fit. Here we consider the following additional complexities to the model structure:

- Model B: $\boldsymbol{X}_i^*(t_{ij}) = \mathbf{1}_{n_i}$ with correlation $\boldsymbol{C}_i$ described by a CAR1 model
- Model C: $\boldsymbol{X}_i^*(t_{ij}) = (\mathbf{1}_{n_i}, \boldsymbol{t}_i)$ without within-group correlation
- Model D: $\boldsymbol{X}_i^*(t_{ij}) = (\mathbf{1}_{n_i}, \boldsymbol{t}_i)$ with correlation $\boldsymbol{C}_i$ described by a CAR1 model

Given the training data the log-likelihood estimates for all models are tabulated in Table 7.

Table 7: Comparison of log-likelihood for different models

	Model A	Model B	Model C	Model D
random effect $\boldsymbol{X}_i^*$	$\mathbf{1}_{n_i}$	$\mathbf{1}_{n_i}$	$(\mathbf{1}_{n_i}, \boldsymbol{t}_i)$	$(\mathbf{1}_{n_i}, \boldsymbol{t}_i)$
correlation $\boldsymbol{C}_i$	none	CAR1	none	CAR1
diabetic nephropathy	-17.7	-0.3	-0.8	1.5
glomerulonephritis	-79.1	-55.6	-43.2	-41.0
HKD	45.5	57.5	51.6	58.9
other	12.8	16.5	21.4	22.0
PKD	24.5	34.7	42.8	45.9
pyelonephritis	64.7	67.9	79.0	78.5
renovascular	51.8	61.2	59.5	64.8
unknown	-23.7	-23.7	-20.0	-20.1
single model all diseases	-488.3	-333.2	-325.8	-302.9

Note: For each disease, the fixed effects derived from Model A are used in Models B to D.

The model which maximises the log-likelihood for each disease category and so gives the best fit

to the data is Model D; see Table 7. Model C generally has a higher log-likelihood than Models A and B, except for diabetic nephropathy, HKD and renovascular where Models C and B have very similar log-likelihoods.

We acknowledge that simply comparing log-likelihood values between models is naive and that from a statistical standpoint model comparison requires likelihood ratio tests. However our model choice is more pragmatic than statistical, in that we took into consideration the known structure of the data (i.e. we expected to need intercept and slope random effects) and although we would have preferred to properly consider correlation in the form of a CAR1 model this could not be achieved within the scope of this thesis as explained below in the second bullet point. Given every disease category, for our final model structure we chose the more parsimonious model, Model C, over Model D. This decision was based on the following considerations:

- Given the aforementioned caveats relating to exploring within-group correlations and likelihood ratio tests we note that the log-likelihood for Model C was only marginally less than Model D, but still approximately matches or is better than models A and B.
- We encountered problems which we could not resolve when fitting Model D to many of the bootstrap samples, specifically the R function nlme::lme() for fitting the mixed effects model reported singularity errors. It is possible that Model D was too complex; a full and detailed investigation was beyond the scope of this thesis. In contrast a model fit was possible for all randomly generated bootstrap samples when using Model C.
- For a given fixed effect parameter all 95% confidence intervals overlapped when comparing these intervals between Models A, B, C and D; note that parameters were selected using Model A. This comparison held true for all fixed effect parameters in all our disease categories. We conclude that these parameter distributions are not statistically different between the models. This means the choice of random effect does not dramatically alter the distribution of the fixed effect parameter values, therefore from this perspective Models A to D are all viable choices.

Step 5. Lastly, given the final choice of model, Model C, we repeated the procedure stated in bullet points of *Step 2.* above. There was little change in the selected fixed effect terms when fitting Model C compared with A.

Finally, for the remainder of this thesis we use Model C where the random effects are accounted for by intercept and slope (followupTime) terms.

6 Diagnostics

In this chapter we verify the robustness of our models, and hence results, by subjecting them to diagnostic tests which predominantly aim to check the linear mixed model assumptions.

6.1 LME Model assumptions

Before reporting results we check the models for each disease are robust and adhere to the basic LME model assumptions, which are:

1. Within-group errors $\boldsymbol{\epsilon}_i$ are independent and identically normally distributed, with zero mean and constant variance.
2. Random effects are normally distributed, with mean zero and covariance matrix Ψ, and are also independent of within-group errors.

We mostly follow diagnostic tests recommended by (66) (e.g. Chapter 4.3) so predominantly concentrate on displaying diagnostic information in plots since, as (66) points out, they are rarely contradicted by hypothesis tests.

6.2 Tests using validation data

When considering model fits to the validation data we note the limitations raised in Section 8.3. In particular we acknowledge that the tests detailed below offer no more than a weak test of overfitting.

On a parameter-by-parameter basis, we compared model estimates fitted using training data with those fitted using validation data; in particular we examined fixed effect parameter estimates, standard errors and confidence intervals. All estimates were very similar, with almost all (training and validation data) confidence intervals overlapping for each parameter.

We examined the residuals of each disease model fit using diagnostic plots (not shown) similar to those in Section 6.6, Figures 8-16. When fitting the models with either the training or validation data we did not find any concerning autocorrelations or deviations from normality. Moreover the plots displayed very similar characteristics for each dataset, these characteristics can also be seen in Figures 8-16 which were created when fitting models to the full dataset.

In summary, there was no concerning evidence of overfitting to the training data For the remainder of this thesis we use the full dataset unless otherwise stated.

6.3 Examination of confidence intervals

Very wide or indeterminate confidence intervals for the LME model parameters indicate numerical instability, consequently the fitted model could not be expected to reliably describe the data. Tables 8 and 9 confirm the confidence intervals for the random effects variance-covariance parameters and σ, give no cause for concern. Figure 5 displays correlation values from Table 8, it clearly shows PKD has a relatively high correlation. The fixed effect confidence intervals, not shown, were also acceptable for each model.

We note in Table 8 that the correlation between random effects is computed from the variance-covariance matrix $\boldsymbol{S}$ i.e. correlation matrix $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\,\boldsymbol{S}\,\boldsymbol{D}^{-1/2}$ where $\boldsymbol{D} = \text{diag}(\boldsymbol{S})$ and the elements of $\boldsymbol{D}^{1/2}$ are standard deviations.

Table 8: 95% confidence intervals for random effects variance-covariance parameters

random effects	lower CI	estimate	upper CI
diabetic nephropathy			
sd(Intercept)	0.284	0.318	0.356
sd(followupTime)	0.043	0.056	0.074
cor(Intercept,followupTime)	-0.174	0.065	0.296
glomerulonephritis			
sd(Intercept)	0.359	0.398	0.441
sd(followupTime)	0.044	0.056	0.071
cor(Intercept,followupTime)	-0.046	0.193	0.412
HKD			
sd(Intercept)	0.299	0.334	0.373
sd(followupTime)	0.028	0.042	0.062
cor(Intercept,followupTime)	-0.219	0.102	0.403
other			
sd(Intercept)	0.310	0.345	0.384
sd(followupTime)	0.029	0.040	0.057
cor(Intercept,followupTime)	-0.179	0.101	0.366
PKD			
sd(Intercept)	0.403	0.472	0.554
sd(followupTime)	0.054	0.077	0.110
cor(Intercept,followupTime)	0.383	0.778	0.932
pyelonephritis			

Table 8: 95% confidence intervals for random effects variance-covariance parameters *(continued)*

random effects	lower CI	estimate	upper CI
sd(Intercept)	0.298	0.350	0.410
sd(followupTime)	0.023	0.033	0.046
cor(Intercept,followupTime)	-0.267	-0.017	0.236
renovascular disease			
sd(Intercept)	0.295	0.342	0.396
sd(followupTime)	0.034	0.048	0.066
cor(Intercept,followupTime)	-0.059	0.331	0.633
unknown			
sd(Intercept)	0.280	0.314	0.352
sd(followupTime)	0.028	0.041	0.061
cor(Intercept,followupTime)	-0.476	-0.169	0.174
single model all diseases			
sd(Intercept)	0.350	0.365	0.380
sd(followupTime)	0.048	0.053	0.058
cor(Intercept,followupTime)	-0.047	0.012	0.071

Table 9: 95% confidence intervals for within-group standard deviation for parameter σ

σ	lower CI	estimate	upper CI
diabetic nephropathy	0.142	0.152	0.164
glomerulonephritis	0.155	0.165	0.175
HKD	0.126	0.136	0.147
other	0.155	0.165	0.177
PKD	0.103	0.116	0.130
pyelonephritis	0.113	0.123	0.135
renovascular	0.125	0.137	0.149
unknown	0.148	0.160	0.173
single model all diseases	0.158	0.162	0.166

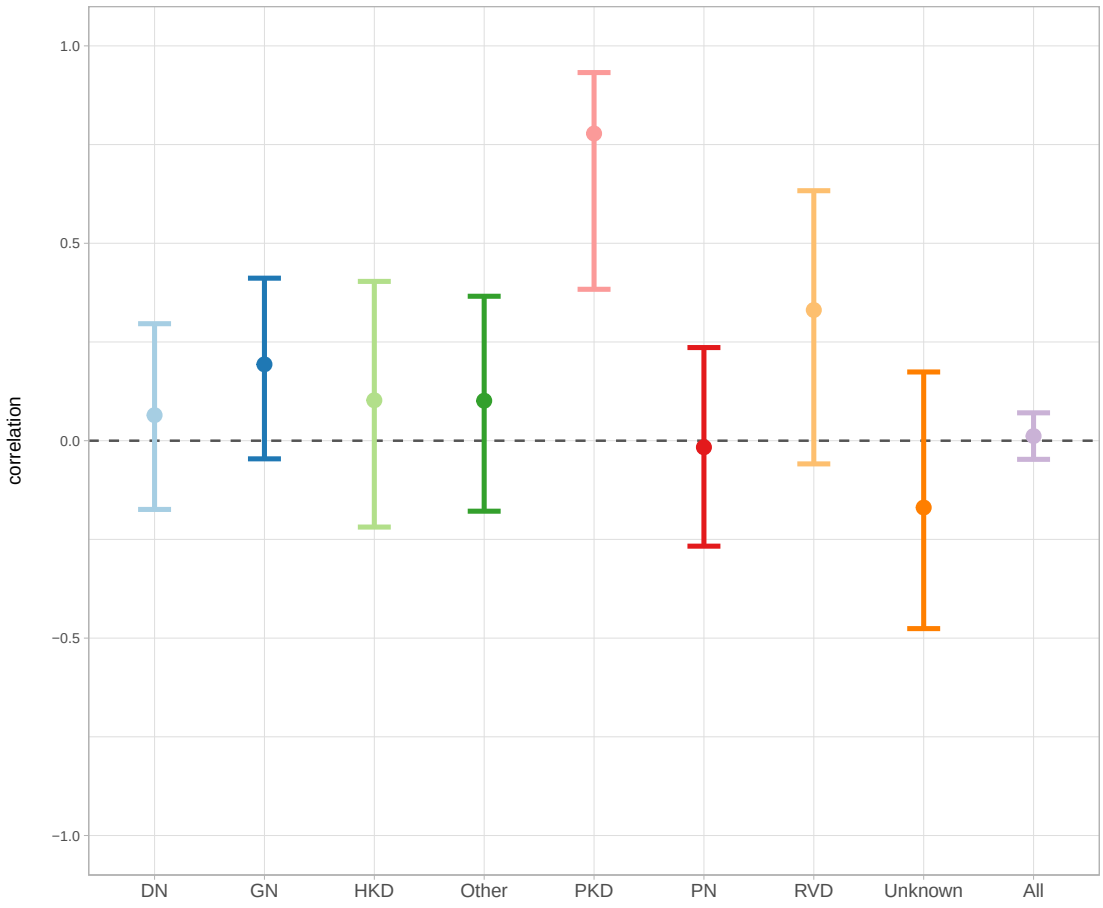

Figure 5: Correlation between intercept and slope random effects

6.4 Observed versus fitted values

For each disease category we show the relationship between the model fitted values and observed values i.e. log(eGFR). Figure 6 depicts observed values plotted against fitted values obtained using a model with fixed effects only (excluding random effects). This gives a summary of the overall quality of the model fixed effects; in all plots there is a reasonable degree of correlation. When using the full model the fitted values include both fixed and random effects, in Figure 7 we observe a marked increase in correlation across all disease categories. This provides evidence that random effects are needed in our models to help explain log(eGFR). For example, given the category 'single model all diseases' the correlation between observed and fitted values without random effects is 0.73, whereas when random effects are included the correlation increases to 0.97.

Given Figure 6 we observe, that compared with the other diseases, PKD has a noticeably wider spread of values. We attribute this to the fixed effects describing the data less well. The dominant determinant for the progression of kidney disease in PKD patients is typically the extent and rate of growth of cysts in the kidneys. Our data does not contain information relating to kidney cysts, therefore this factor cannot be included in the PKD model fixed effects. This possibly explains why we observe a wider spread in values in Figure 6. This wide spread of values for PKD is not seen in Figure 7 hence the inclusion of the random effects accounts for the additional variability not accounted for by the fixed effects.

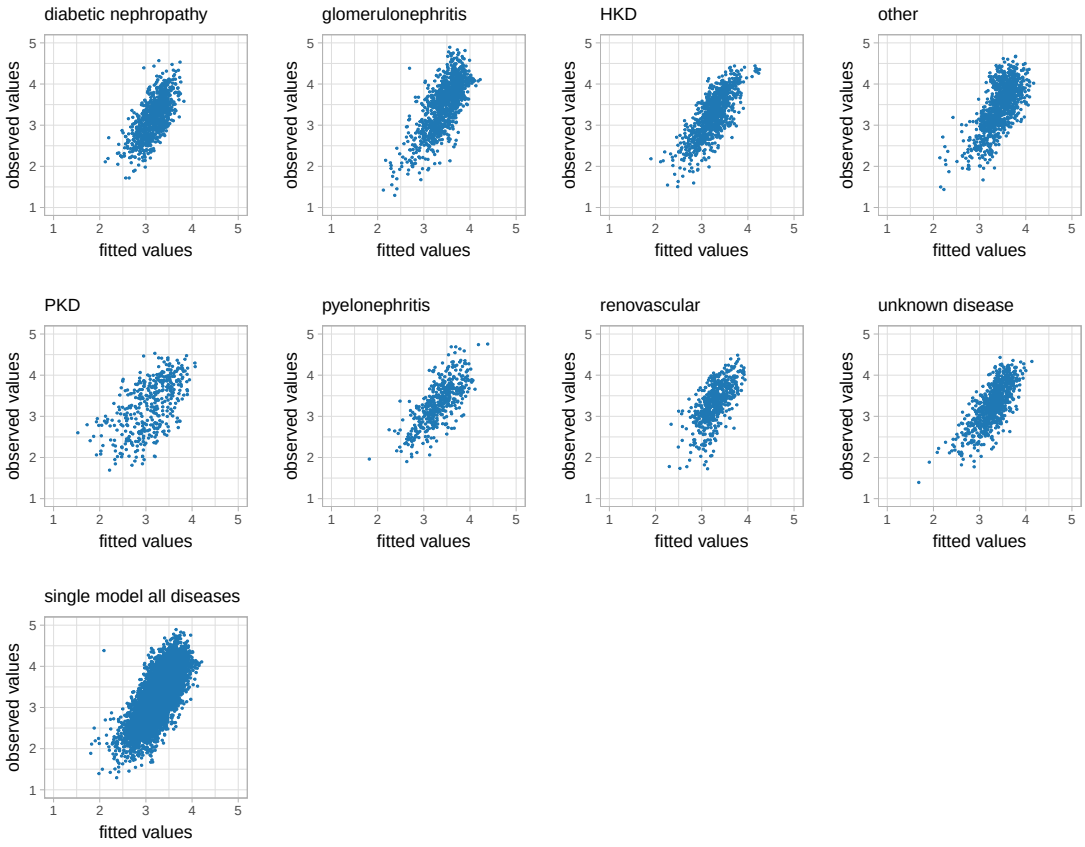

Figure 6: Observed values plotted against fitted values obtained using a model with fixed effects only

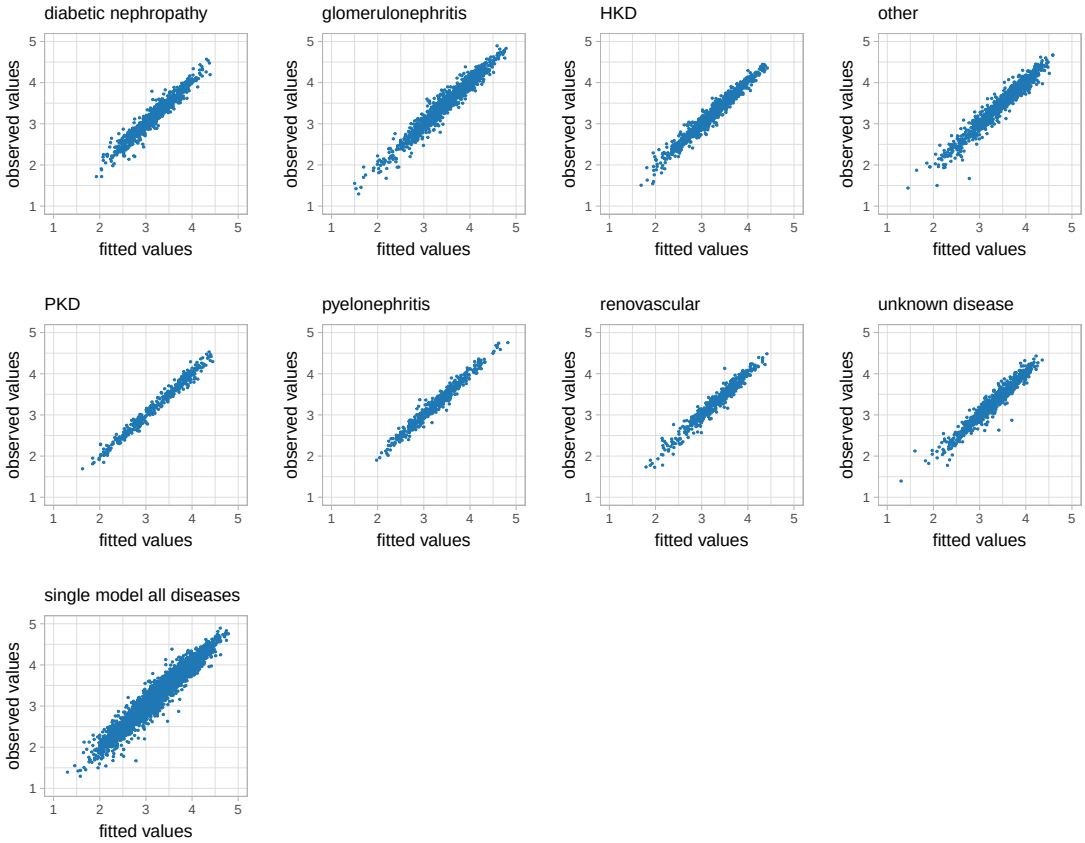

Figure 7: Observed values plotted against fitted values obtained using full model with fixed and random effects

6.5 Assessment of residual distributional assumptions

Standardised (or Pearson) residuals are found by subtracting the estimated fitted value vector from the outcome vector, then dividing through by the corresponding estimated within-group standard errors. Fitted values are obtained by adding the estimated contributions from both fixed and random effects vectors. We expect the standardised residuals to follow a standard normal distribution.

In Figures 8-16 we assess, for each disease, the normality assumptions of the residuals using a panel of four plots:

- *Left plot* - standardised (or Pearson) residuals against fitted values. From these plots we report that the residuals in our LME models are reasonable given the within-group error assumptions: the residuals are symmetrically distributed around zero with approximately constant variance.
- *Left middle plot* - qq-plot with standardised residual quantiles against theoretical quantiles. These plots confirm that our models have residuals which are plausible under the assumption of normality. However outside of about -2 to 2 quantiles there are more than expected extreme positive and negative residuals hence our distributions have long tails. Clearly our models do not adequately explain the extremes however this is not a significant issue given our objective is to identify fixed effect parameters and not to make predictions (note that with our model's predictions based on the extremes would be poor).
- *Right middle plot* - cumulative probability for both the standard normal distribution (dotted black) and standardised residual (solid blue). These plots confirm that our residuals do not indicate any significant violations of the normality assumption.
- *Right plot* - empirical autocorrelation function for standardised residuals where lag is the difference between follow-up years and the shaded area is the 95% CI. These plots show that there exists some autocorrelation which is not accounted for by the LME model however we consider this amount of autocorrelation to be acceptable. The large correlations at large lags are most likely due to the small numbers patients followed-up over many years e.g. PKD has less than 20 patients beyond four follow-up years. For reference, the number of patients at each follow-up year are denoted by 'n=...' in Figures 63-71.

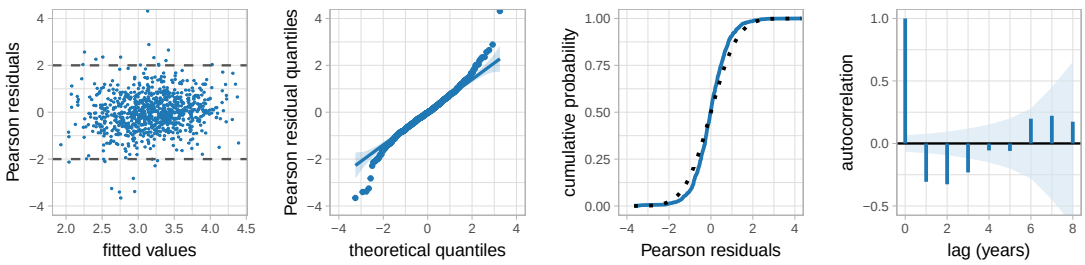

Figure 8: Residuals: disease diabetic nephropathy

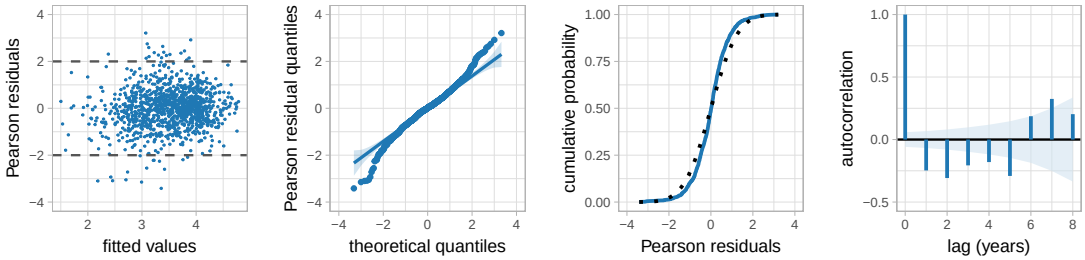

Figure 9: Residuals: disease glomerulonephritis

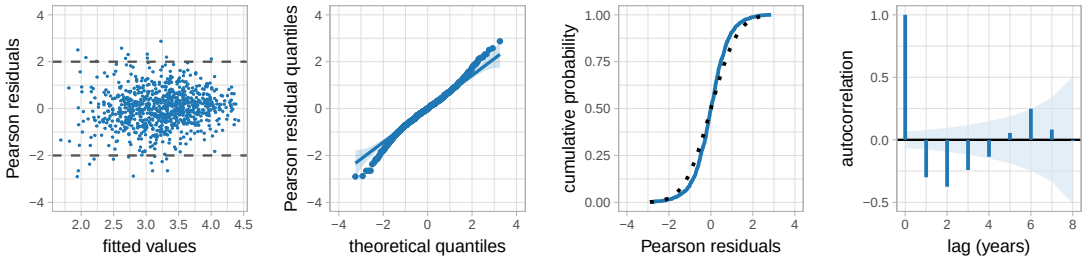

Figure 10: Residuals: disease HKD

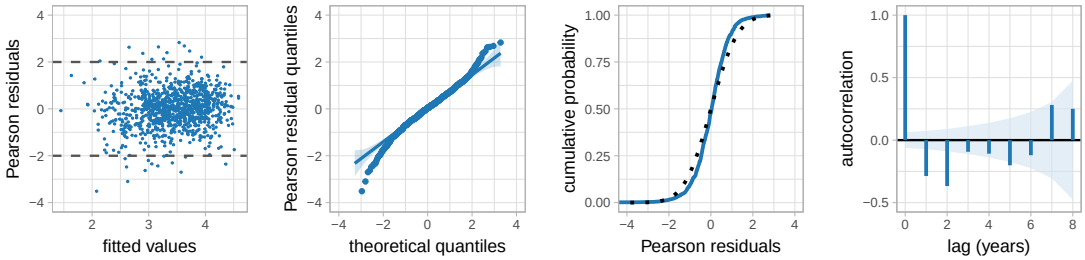

Figure 11: Residuals: disease other

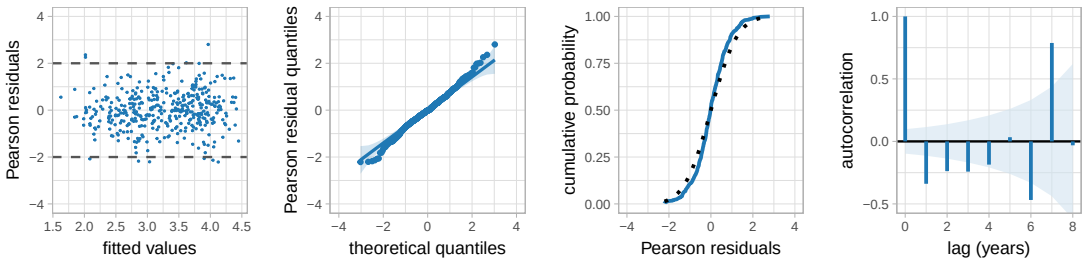

Figure 12: Residuals: disease PKD

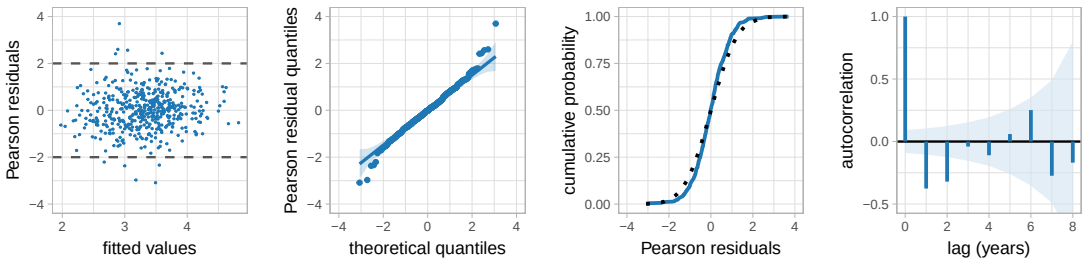

Figure 13: Residuals: disease pyelonephritis

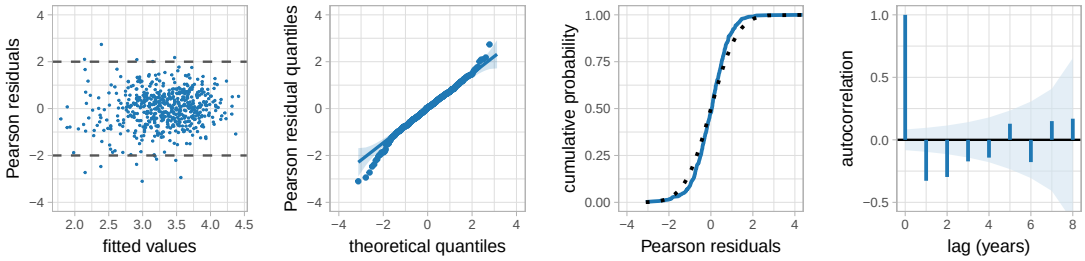

Figure 14: Residuals: disease renovascular

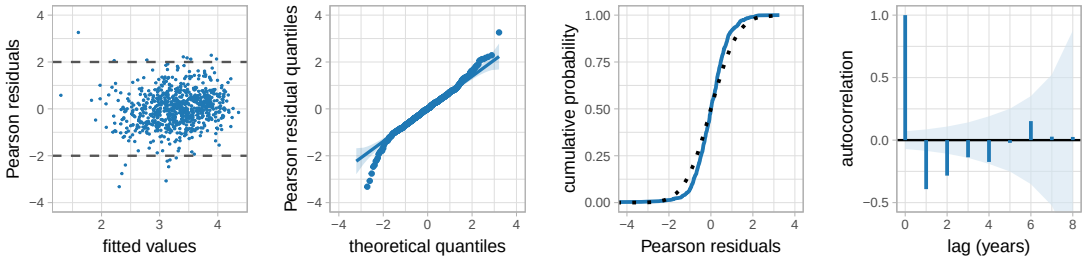

Figure 15: Residuals: disease unknown

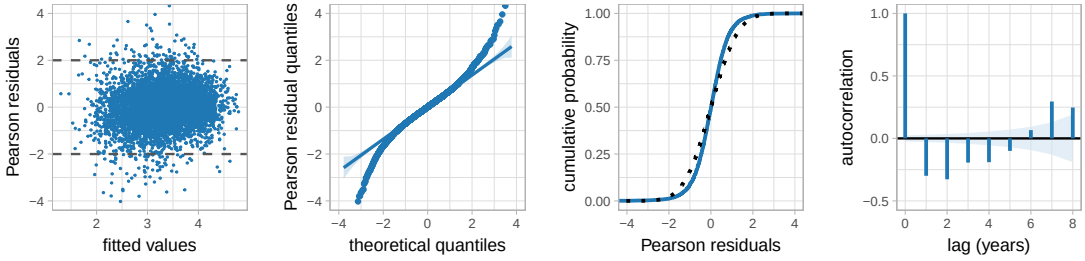

Figure 16: Residuals - single model all diseases

Appendix A.5, Figures 63-71, show there is no systematic trend in the mean of the residuals over time and furthermore the majority of 95% confidence intervals cover zero.

Despite the aforementioned weaknesses we conclude for each disease category that the residuals do not show any concerning deviation from normality. With reference to Section 6.1 we conclude that assumption 1 is sufficiently true; i.e. errors are independent and identically normally distributed, with zero mean and constant variance.

Removing outliers

When we present the results in Chapter 7 we will exclude the following outliers:

- given figure 8 the record with the residual value of 4.1 will be removed from diabetic nephropathy
- given figure 13 the record with the residual value of 3.3 will be removed from pyelonephritis
- given figure 15 the record with the residual value of 3.2 will be removed from disease unknown

These outliers were far from any other observations within their disease category and therefore might skew the parameter estimation. We took this approach since we are unable to determine if the outliers were due to: a) natural variability not accounted for by our model; b) measurement error; c) data recording error; or d) sub-optimal imputation. Future work should carefully investigate outliers as they may be medically informative if caused by unusual but interesting biological mechanisms not accounted for within our models. However we note that the removal of the aforementioned outliers made negligible difference to our parameter estimates.

6.6 Assessment of random effect distributional assumptions

Given each disease category the mean value of each estimated random effect vector is, as required, approximately zero; range -1.2e-14 to 1.3e-13.

Figures 17 and 18 respectively show the qq-plots of the estimated random effects for slope and intercept, it can be seen that the assumption of marginal normality is plausible although the distributions for the slope term deviate from normality beyond about 1 standard deviation.

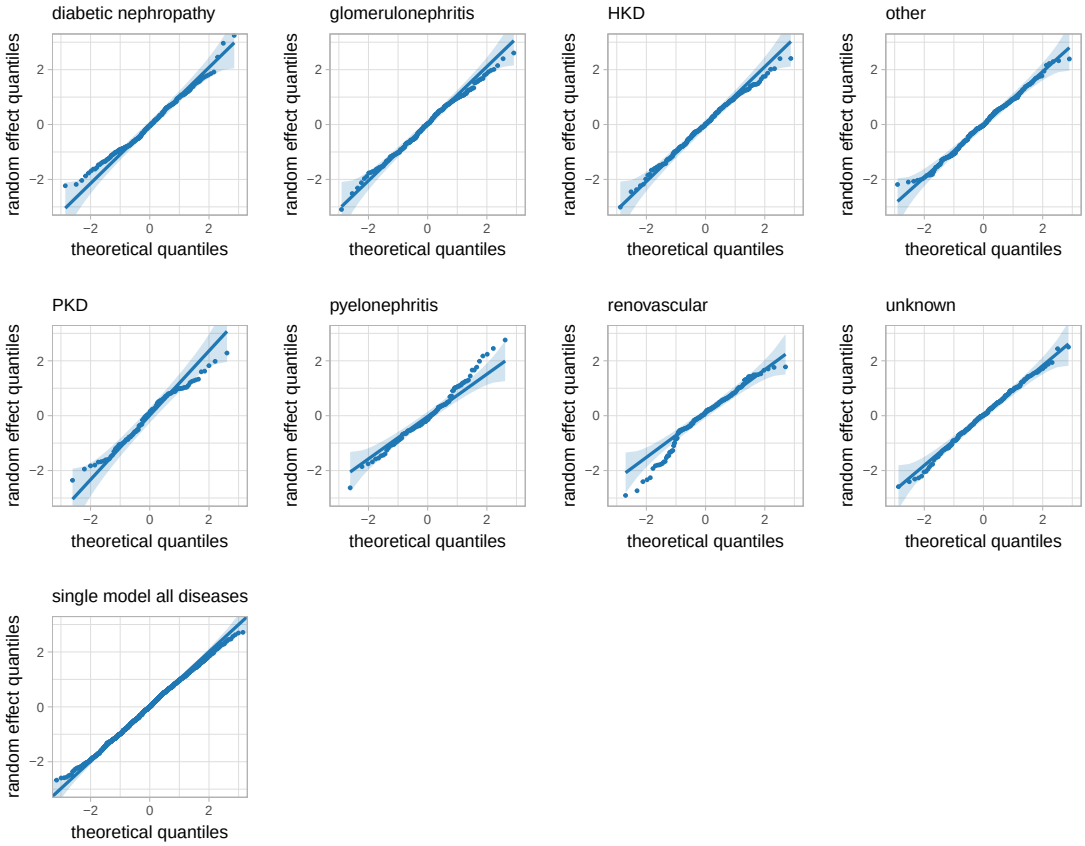

Figure 17: qq-plot for standardised random effect intercept term

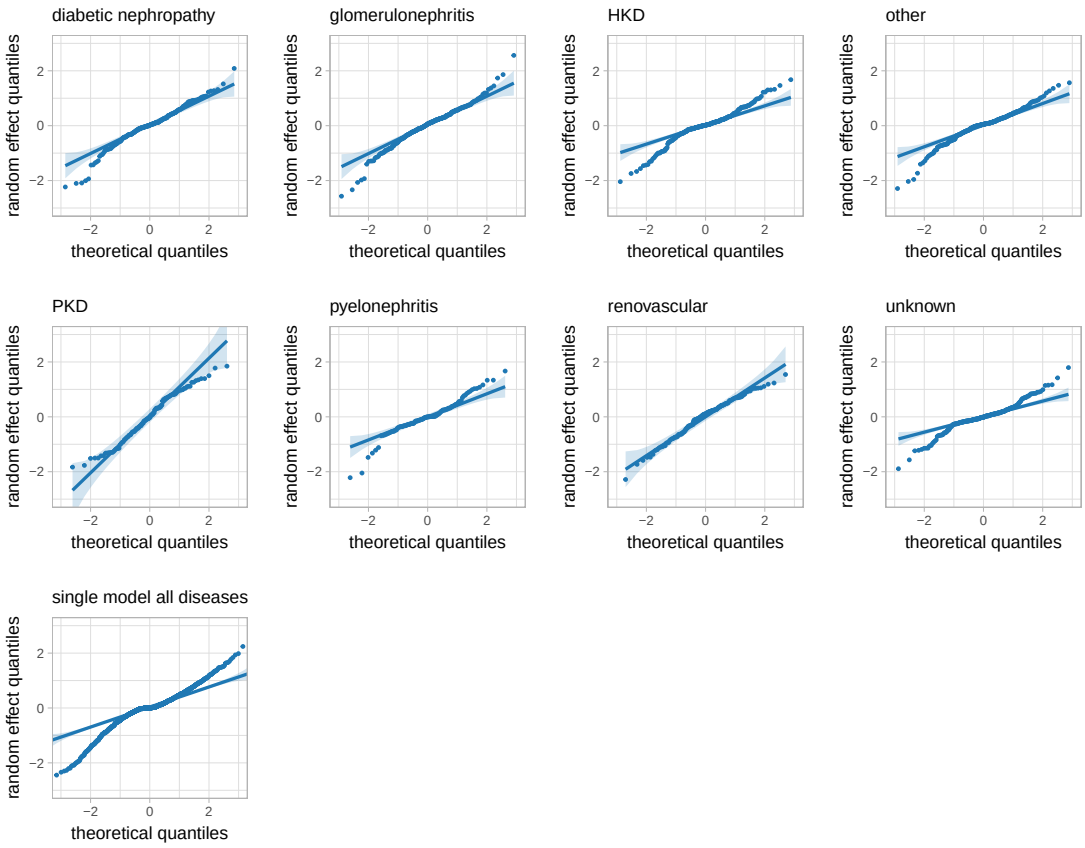

Figure 18: qq-plot for standardised random effect slope term

Despite the aforementioned weaknesses we conclude for each disease category that the random effects do not show any concerning deviation from normality, with reference to Section 6.1 we conclude that assumption 2 is plausible; i.e. random effects are normally distributed, with mean zero.

As an aside we investigate the extent to which there are correlations between the estimated random effects terms i.e. intercept and slope. We expect that there may be some correlations, although in terms of our model fitting this is not a concern. For most diseases there is no significant correlation, the exception is PKD which is highly correlated.

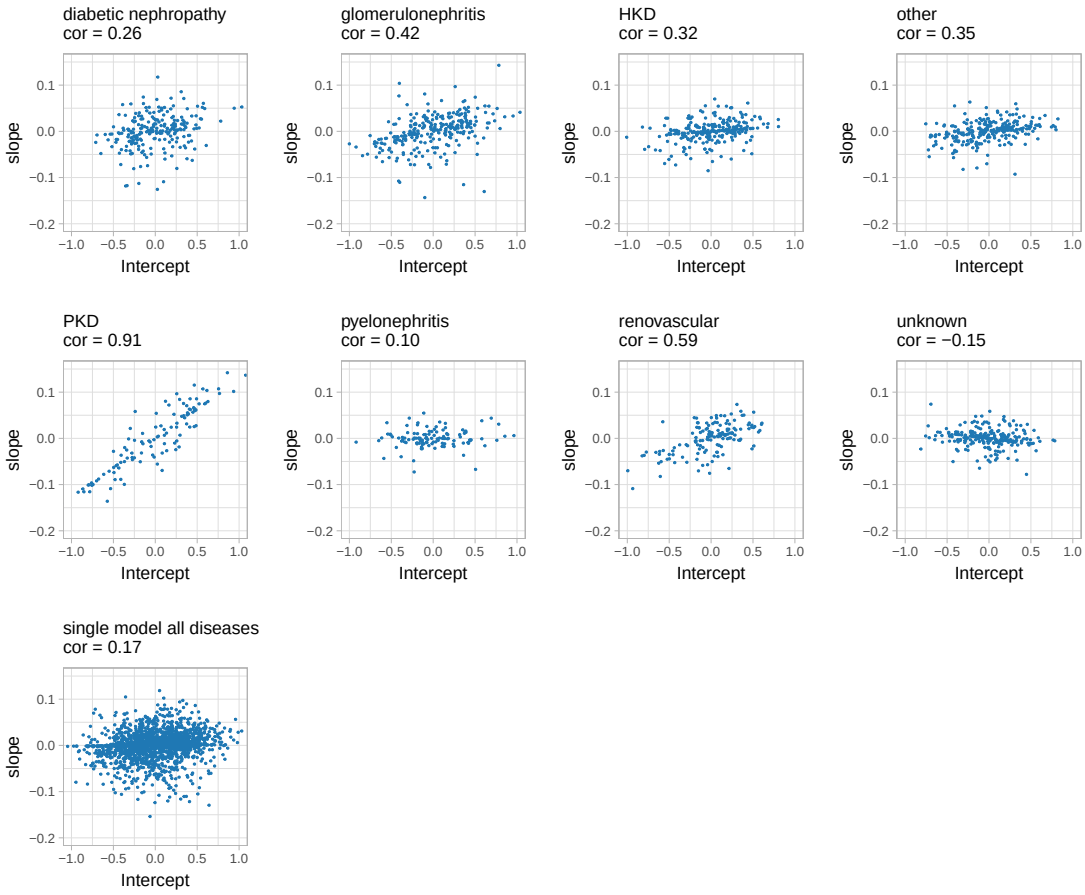

Figure 19: Estimated random effects plotted against each other

6.7 Robustness of fixed effect parameters and conclusions relating to diagnostic results

In this thesis our primary interest is in fixed effects therefore we consider how they are influenced by the choice of random effects. We rewrite Equation 2 such that

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i^*\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^* \end{aligned} \tag{25}$$

where $\boldsymbol{\epsilon}_i^* = \boldsymbol{X}_i^*\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Psi})$. Both $\boldsymbol{X}_i^*\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are independently distributed as multivariate normal vectors hence their sum $\boldsymbol{\epsilon}_i^*$ is an independently distributed multivariate normal vector with mean zero and variance-covariance matrix $\sigma^2\boldsymbol{\Sigma}_i = \boldsymbol{X}_i^*\boldsymbol{\Psi}(\boldsymbol{X}_i^*)^T + \sigma^2\boldsymbol{I}$. Consequently $\boldsymbol{Y}_i$ are independent multivariate normal random vectors with variance-covariance matrix $\sigma^2\boldsymbol{\Sigma}_i$ and mean $\boldsymbol{X}_i\boldsymbol{\beta}$ i.e. $\boldsymbol{Y}_i|\boldsymbol{X}_i \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}_i)$. The density is

$$P(\boldsymbol{Y}_i|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n_i}{2}} \exp\left(-(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})2^{-1}\sigma^{-2}\right)|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \tag{26}$$

where $|\boldsymbol{\Sigma}_i|$ is the determinant of $\boldsymbol{\Sigma}_i$. The log-likelihood is

$$l(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = C + \sum_{i=1}^{M}((\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}) \tag{27}$$

and we note, using matrix algebra, that

$$(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{Y}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Y}_i - 2\boldsymbol{\beta}^T\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Y}_i + \boldsymbol{\beta}^T\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\boldsymbol{\beta}. \tag{28}$$

Differentiating the log-likelihood with respect to $\boldsymbol{\beta}$, equating to zero, and evaluating at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ gives the maximum likelihood estimate for this parameter;

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{M}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Y}_i. \tag{29}$$

Using the law of total expectation, i.e. $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$, and given $\mathbb{E}(\boldsymbol{\epsilon}_i^*|\boldsymbol{X}_i) = \boldsymbol{0}$, then

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \left(\sum_{i=1}^{M} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{M} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned} \tag{30}$$

It follows that the estimator of fixed effects parameter $\boldsymbol{\beta}$ is unbiased where the only assumption required is $\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{X}_i^*) = \boldsymbol{X}_i\boldsymbol{\beta}$. This result shows that the LME model gives unbiased estimates of $\hat{\boldsymbol{\beta}}$ even if the random effects and residual assumptions are violated. For example, our intercept-and-slope random effects assume linear slopes over time but if the patient's have longitudinal trajectories which are actually non-linear we will still obtain unbiased estimates of $\hat{\boldsymbol{\beta}}$ despite the departure from normality as depicted in Figure 18; here we assume reasonable fixed effects terms. Even if the random effects and/or residuals, are not normally distributed, and/or the variances are not constant, we still obtain unbiased estimates of $\hat{\boldsymbol{\beta}}$ but we should be careful when drawing inferences; estimating standard errors, and calculating confidence intervals and p-values.

If one or more of the LME assumptions are violated our inferences relating to the fixed effects should still be valid since, as discussed above these models are robust to such violations. We therefore conclude that our LME model diagnostics do not reveal any deviations from normality or homoscedasticity which are sufficient to cause us concern regarding the robustness of the fixed effect parameters estimates.

7 Results

7.1 Introduction

We present our results based on the LME models for each disease category that we fitted in Chapter 5. In Section 7.2 we give an overview of regression parameter estimates and report their details in Section 7.3. These two sections constitute our main findings regarding the key factors affecting kidney disease. As discussed in Section 4.5 all continuous explanatory variables have been standardised except for follow-up time and baseline age. For reference only, the equivalent regression parameter estimates in which variables are not standardised are given in Appendix A.7. In Section 7.4 we report the rates at which the variables change over time; these results are of secondary importance relative to Sections 7.2 and 7.3 .

The regression parameter estimates are difficult to interpret from a clinical perspective because they are relative to the log(eGFR) scale, most clinicians work in terms of eGFR. We therefore report the *relative change* in eGFR induced by a step change in the variable of interest, see Equation 14. A relative change of 5% in eGFR is generally considered to be a clinically significant. Given standardised variables we set $\theta_r^{'} = 1$ therefore the parameter effects are comparable. In this thesis a superscript dash is never used to denote a derivative, here we use a superscript dash, e.g. $\theta_r^{'}$, to denote that the term belongs to the standardised model as described in Section 4.5.

When we report p-value estimates for regression parameter values the null hypothesis is that the parameter is zero valued i.e. the parameter has no effect on the outcome. We choose to reject the null hypothesis at the 5% significance level, hence with this interpretation a p-value < 0.05 indicates changes in the predictor are associated with meaningful changes in the outcome.

For each disease model the details of the fixed effect parameter estimates are given in Tables 10 to 20. Additionally Figures 20 to 37 summarise the relative change in eGFR for $\theta_r^{'} = 1$ and indicate the clinically significant level of a 5% change in eGFR.

When reporting the regression parameters for a given disease we split them into two categories as follows:

- *average effects* - describe the average behaviour of the population. These effects related to explanatory variables which do not have an interaction with time. Parameters with positive values indicate a higher level in eGFR (less severe kidney disease) whereas negative values indicate a lower level in eGFR (more severe kidney disease).
- *temporal effects* - describe the explicit time dependent behaviour of the population. These effects relate to explanatory variables which have an interaction with follow-up time, that is $X_i^{(r)}(t_{i,j})\ t_{i,j}$ or equivalently we denote such terms as *explanatoryVariable : followupTime* (this is the notation used by R-package nlme). As shown in Section 7.4.1 the eGFR is

on average decreasing over time, i.e. it has a negative slope over time, for each disease category. It follows that parameters with positive values reduce the gradient of the slope of eGFR, that is the slope is less negative (more shallow), which suggests a less rapid decline in kidney function. Conversely, negative parameter values indicate an increase in the slope of eGFR, that is the slope is more negative (steeper), which suggests a more rapid decline in kidney function. Mathematically this can be seen by considering the linear model $y(x,t) = ax + bt + cxt = ax + (b + cx)t$ where the slope with respect to t is $(b + cx)$. It follows that if b is negative, then for positive c an increase in x leads to a slope which is less steep. Conversely for negative b and c, an increase in x results in a slope that is steeper.

The clinicians advised that, within the following three groups, the variables are clinically strongly associated:

- med.VitaminD, CC, PO, PTH,
- med.ACE.ARB, numberAntihypertensives, DBP, PP and Pu
- med.iron (iron taken orally), med.ParenteralIron, med.Epo and Hb

If one or more variables within a given group are selected by our model selection procedure (see Section 5.4) then all variables from that group will be used when reporting the results. The reasoning behind these strong associations is that within each of these groups the medications change the levels of the biochemical markers, and where applicable the blood pressure markers (DBP, PP).

7.2 Overview

In Tables 10 and 11 we present a summary of the fixed effect regression parameters for each disease model. In these tables the 'single model all diseases' column is the model which encompasses all disease categories including the obstruction category. If a parameter is present in a given model it is denoted by either star(s) or tilde. One or more stars indicate that we reject the null hypothesis, hence the parameter is statistically significant. Three stars indicate we reject the null hypothesis at the significance level of 0.001, two stars at a level of 0.001-0.01 and one star at a level of 0.01-0.05. Tilde indicates although we do not reject the null hypothesis we still include the parameter in the model as it has a small effect. The plus and minus signs inside the brackets indicate the sign of the estimated parameter, for example (-)*** denotes the given parameter estimate has a negative value and is statistically significant at the 0.001 level.

Key messages

As expected there is much variation across disease categories. Purely in terms of parameter p-values for the *average effects*, given in Table 10, we observe the following:

- In every disease model parameters baseline age, vitamin D, Hb, PO and PTH are typically highly statistically significant.
- Medications play a stronger role than comorbidities.
- Baseline lifestyle parameters (living alone, occupation, smoking and weekly alcohol intake) do not play as strong a role as we might have anticipated.
- Physical attributes such as BMI, sex and ethnicity generally have a very weak effect.
- Each of the CC, DBP, Hb, PO, PP, Pu, PTH, total Cholesterol and total CO2 factors are statistically significant in at least one disease category, the exception is CRP which is not statistically significant anywhere.
- medication med.ACE.ARB is statistically significant for diabetic nephropathy and less so for glomerulonephritis but it is not significant for the other diseases.

Considering the *temporal effects*, given in Table 11, we observe:

- The majority of these variables the steepness of their slopes are not statistically significant. For example the comorbidities very weakly influence the progression of disease.
- In some cases follow-up time is significant, especially so for PKD. This most likely indicates there are other risk factors which are not in our model.
- For some diseases, the biomarkers DBP, Hb and PTH have a positive effect on the slope indicating less rapid progression of disease. In contrast PO, Pu, total cholesterol and total CO2 have a negative effect indicating a more rapid progression of disease.

Where anomalies occur in the signs of parameters they are discussed in detail with reference to each primary kidney disease within Section 7.3.

Average effects

Table 10: Average effects - standardised model summary for each disease

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
(Intercept)	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$
age0	$(-)^{**}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{**}$	$(-)^{*}$	$(-)^{***}$	$(-)^{***}$
bodyMassIndex	$(-)^{**}$					$(+)^{\sim}$			
CC	$(-)^{**}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{**}$
comorbidityCancercurrent			$(+)^{\sim}$			$(-)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$
comorbidityCancerprevious			$(-)^{*}$			$(-)^{\sim}$		$(+)^{\sim}$	$(-)^{\sim}$
comorbidityCV1			$(-)^{*}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{*}$
comorbidityCVover 1			$(-)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{*}$		$(-)^{**}$
comorbidityDiabetestype1				$(-)^{\sim}$			$(-)^{\sim}$		
comorbidityDiabetestype2				$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$		
comorbidityGastrointestinal						$(+)^{\sim}$	$(+)^{\sim}$		
comorbidityOther			$(+)^{\sim}$						
CRP		$(+)^{\sim}$		$(+)^{\sim}$	$(-)^{\sim}$		$(+)^{\sim}$		
DBP	$(+)^{*}$	$(-)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{*}$	$(+)^{*}$
disease diabetic nephropathy									$(-)^{**}$
disease glomerulonephritis									$(+)^{\sim}$
disease HKD									$(-)^{\sim}$
disease obstruction									$(-)^{***}$

Table 10: Average effects - standardised model summary for each disease *(continued)*

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
disease polycystic kidney disease									$(-)^{***}$
disease pyelonephritis									$(-)^{**}$
disease renovascular disease									$(-)^{\sim}$
disease unknown									$(-)^{\sim}$
ethnicitynonWhite			$(+)^{\sim}$						
familyHistoryIHD0					$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$		
Hb	$(+)^{\sim}$	$(+)^{**}$	$(+)^{***}$	$(+)^{***}$	$(+)^{\sim}$	$(+)^{*}$	$(+)^{**}$	$(+)^{***}$	$(+)^{***}$
med.ACE.ARB	$(+)^{***}$	$(+)^{**}$	$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{**}$
med.AlphaBlockers		$(-)^{\sim}$				$(-)^{\sim}$	$(+)^{\sim}$		$(-)^{\sim}$
med.BetaBlockers	$(+)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$				$(+)^{\sim}$	
med.CCBs		$(-)^{**}$	$(-)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$	$(-)^{*}$
med.Diuretics					$(+)^{\sim}$		$(-)^{*}$		$(-)^{**}$
med.Epo	$(-)^{**}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{*}$	$(-)^{***}$
med.Iron	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$
med.Other									$(-)^{*}$
med.ParenteralIron	$(-)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$	$(+)^{\sim}$	$(+)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$
med.VitaminD	$(-)^{***}$	$(-)^{***}$	$(-)^{**}$	$(-)^{***}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{*}$	$(-)^{**}$	$(-)^{***}$
numberAKIepisodes	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$		$(+)^{*}$	$(-)^{\sim}$
numberAntihypertensives	$(-)^{*}$	$(-)^{\sim}$	$(-)^{**}$		$(-)^{\sim}$	$(-)^{**}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$

Table 10: Average effects - standardised model summary for each disease

(continued)

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
numberClinicVisits			$(-)^{\sim}$	$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$			$(-)^{\sim}$
occupation0ManagerialProfessional					$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$		
occupation0Intermediate					$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$		
occupation0NeverWorkedUnemployed					$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$		
PO	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{*}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$
PP	$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$		$(+)^{*}$	$(-)^{\sim}$	$(+)^{**}$	$(+)^{\sim}$	$(+)^{*}$
PTH	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{**}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{***}$	$(-)^{***}$
Pu	$(+)^{\sim}$	$(+)^{**}$	$(-)^{\sim}$		$(-)^{*}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$	$(+)^{\sim}$
sexfemale						$(-)^{\sim}$			
smokingStatus0active			$(-)^{\sim}$			$(-)^{\sim}$	$(-)^{\sim}$		
smokingStatus0ex-smoker			$(-)^{\sim}$			$(+)^{\sim}$	$(-)^{\sim}$		
totalCholesterol	$(+)^{\sim}$			$(+)^{\sim}$			$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{*}$
totalCO2		$(+)^{***}$	$(+)^{**}$	$(+)^{***}$	$(+)^{*}$	$(+)^{**}$		$(+)^{**}$	$(+)^{***}$
weeklyAlcohol01 to 14			$(-)^{*}$			$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	
weeklyAlcohol0over 14			$(-)^{\sim}$			$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$	

Note:

regression parameter sign: positive (+); negative (-)

p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *; >0.05 ∼

[1] 'all' denotes 'single model all diseases'

Temporal effects

Table 11: Temporal effects - standardised model summary for each disease

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
bodyMassIndex:followupTime	(+)$^{\sim}$					(+)$^{\sim}$			
CC:followupTime	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$
comorbidityCancercurrent:followupTime			(+)$^{\sim}$			(+)$^{\sim}$		(+)$^{\sim}$	(−)$^{\sim}$
comorbidityCancerprevious:followupTime			(+)$^{\sim}$			(+)$^{\sim}$		(−)$^{\sim}$	(+)$^{\sim}$
comorbidityCV1:followupTime			(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$		(+)$^{\sim}$
comorbidityCVover 1:followupTime			(+)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$		(+)$^{\sim}$
comorbidityDiabetestype1:followupTime				(−)$^{\sim}$			(+)$^{\sim}$		
comorbidityDiabetestype2:followupTime				(+)$^{\sim}$		(+)$^{\sim}$	(+)$^{\sim}$		
comorbidityGastrointestinal:followupTime						(+)$^{\sim}$	(−)$^{\sim}$		
comorbidityOther:followupTime			(−)$^{\sim}$						
CRP:followupTime		(+)$^{\sim}$		(−)$^{\sim}$	(+)$^{\sim}$		(−)$^{\sim}$		
DBP:followupTime	(−)$^{\sim}$	(+)*	(+)$^{\sim}$		(−)$^{\sim}$	(−)$^{\sim}$	(+)*	(−)$^{\sim}$	(+)$^{\sim}$
followupTime	(−)**	(−)$^{\sim}$	(−)*	(−)$^{\sim}$	(−)***	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)***
Hb:followupTime	(+)**	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)*	(+)*	(+)$^{\sim}$	(+)$^{\sim}$	(+)**
med.ACE.ARB:followupTime	(−)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$		(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$
med.AlphaBlockers:followupTime		(−)$^{\sim}$				(+)$^{\sim}$	(+)**		(−)$^{\sim}$
med.BetaBlockers:followupTime	(−)$^{\sim}$	(−)$^{\sim}$		(−)$^{\sim}$				(−)$^{\sim}$	
med.CCBs:followupTime		(+)$^{\sim}$	(+)$^{\sim}$		(−)$^{\sim}$	(−)$^{\sim}$		(−)$^{\sim}$	(−)$^{\sim}$

Table 11: Temporal effects - standardised model summary for each disease

(continued)

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
med.Diuretics:followupTime					(−)$^{\sim}$		(+)$^{\sim}$		(+)$^{\sim}$
med.Epo:followupTime	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$
med.Iron:followupTime	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)*
med.Other:followupTime									(+)$^{\sim}$
med.ParenteralIron:followupTime	(+)**	(−)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$
med.VitaminD:followupTime	(+)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$
numberAKIepisodes:followupTime	(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$		(+)$^{\sim}$		(−)$^{\sim}$	(−)$^{\sim}$
numberAntihypertensives:followupTime	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$		(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(+)*	(−)$^{\sim}$
numberClinicVisits:followupTime			(−)***	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$			(−)$^{\sim}$
PO:followupTime	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(−)**	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)***	(−)**
PP:followupTime	(−)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$		(+)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$
PTH:followupTime	(+)$^{\sim}$	(+)***	(−)$^{\sim}$	(+)**	(+)$^{\sim}$	(+)$^{\sim}$	(+)$^{\sim}$	(+)***	(+)***
Pu:followupTime	(−)***	(−)***	(−)***		(+)$^{\sim}$	(−)***	(−)$^{\sim}$	(−)$^{\sim}$	(−)***
totalCholesterol:followupTime	(−)$^{\sim}$			(−)$^{\sim}$			(−)**	(+)$^{\sim}$	(−)$^{\sim}$
totalCO2:followupTime		(−)$^{\sim}$	(−)$^{\sim}$	(+)$^{\sim}$	(−)$^{\sim}$	(−)*		(−)$^{\sim}$	(−)***

Note:

regression parameter sign: positive (+); negative (-)

p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *; >0.05 ∼

[1] 'all' denotes 'single model all diseases'

7.3 Detailed Estimates of regression parameters

For each disease category, the figures in this section show the *relative change* in eGFR, these values are computed using Equation 14. The tables also report regression parameter estimates, standard errors and the proportion of bootstraps in which each variable was selected.

7.3.1 Diabetic nephropathy

Average effects

The key *average effects* are:

- Lower levels of eGFR are associated with having EPO treatment, vitamin D supplements, PO and PTH. It is known that iron levels drop as kidney disease worsens, so patients requiring EPO treatment have lower levels of eGFR. Similarly vitamin D drops as kidney function worsens, so it is reasonable that these patients require vitamin D supplements. Poor kidney function can result in higher levels PO and PTH and hence lower levels of eGFR.
- Lower levels of CC are associated with lower levels of eGFR and therefore indicate poorer kidney function; this is medically plausible.
- An older age at baseline and higher body mass index are both associated with lower levels of eGFR. Note that increased body mass index is associated with type 2 diabetes.
- The ACE inhibitors and ARBs (med.ACE.ARB) are associated with higher levels of eGFR. This suggests that patients taking ACE inhibitors and/or ARBs have better kidney function compared with those who are not taking these drugs. We note that if eGFR drops to a very low value then the patient is taken off these drugs.

PP and DBP are less statistically significant (level 0.01-0.05) then the key results stated above. However we note that from a medical perspective it is expected that PP will increase with ageing, worsening CKD and increasing cardiovascular disease. Our PP results are consistent with this view. However as PP increases it is generally expected that DBP will fall; our results are counterintuitive in this regard.

If HbA1c is included in the model then its corresponding parameter is positive but it has a significance level greater than 0.05.

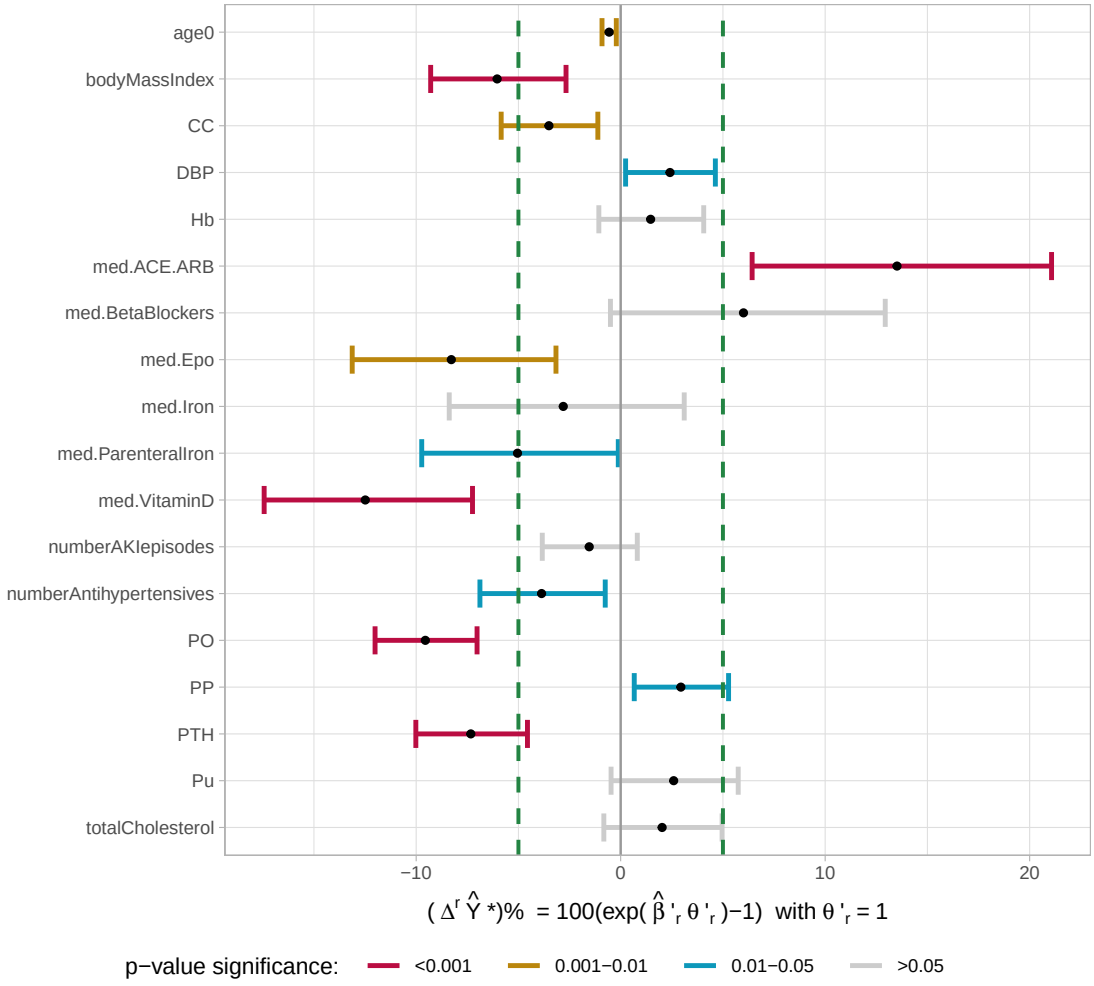

Figure 20: Average effects - relative change in eGFR for standardised model using 95% CIs: diabetic nephropathy

Temporal effects

The key *temporal effects* are:

- Follow-up time is negative. This means that the level of eGFR is falling off over time. This is to be expected. The significance of follow-up time may indicate that there are additional risk factors which are not included in the SKS dataset.
- Hb and parenteral iron are associated with slower disease progression (less negative slope in eGFR). Note that Hb levels can fall as renal disease worsens.
- Pu is associated with a steeper decline in eGFR

If HbA1c is included in the model then its corresponding interaction term HbA1c:followupTime has a parameter which is negative and significant at the 0.01-0.05 level.

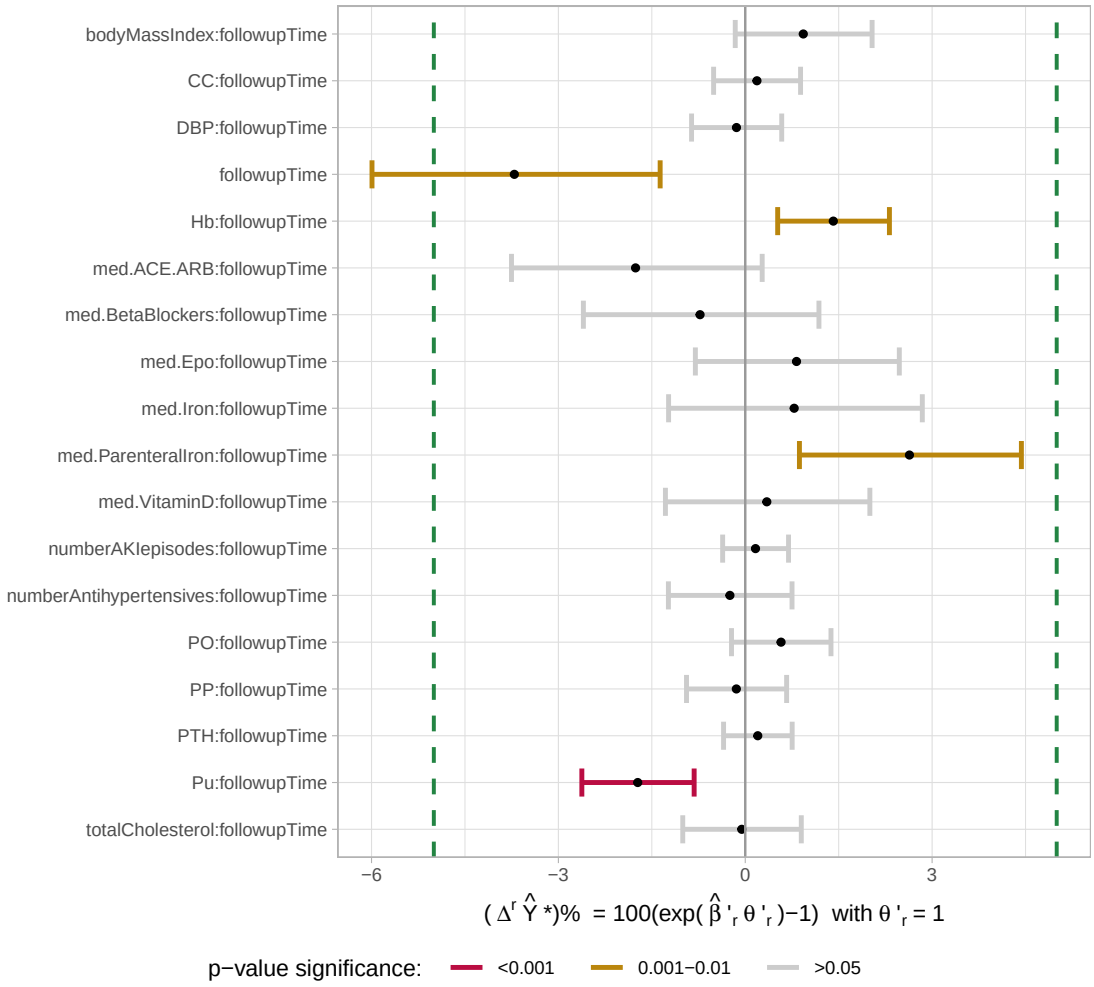

Figure 21: Temporal effects - relative change in eGFR for standardised model using 95% CIs: diabetic nephropathy

Parameter values

Table 12: Standardised model summary for disease diabetic nephropathy

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	3.5808	1.2e-01	0.000	***	
age0	0.98	-0.0057	1.8e-03	0.002	**	-0.56
bodyMassIndex	0.52	-0.0623	1.8e-02	0.001	**	-6.04
bodyMassIndex:followupTime		0.0093	5.7e-03	0.102		0.93
CC	0.76	-0.0357	1.3e-02	0.005	**	-3.51
CC:followupTime		0.0019	3.6e-03	0.605		0.19
DBP	0.88	0.0238	1.1e-02	0.033	*	2.41
DBP:followupTime		-0.0014	3.8e-03	0.709		-0.14
followupTime	1.00	-0.0378	1.2e-02	0.003	**	-3.71
Hb	0.99	0.0145	1.3e-02	0.269		1.46
Hb:followupTime		0.0140	4.6e-03	0.002	**	1.41
med.ACE.ARB	0.95	0.1268	3.4e-02	0.000	***	13.51
med.ACE.ARB:followupTime		-0.0178	1.1e-02	0.096		-1.76
med.BetaBlockers	0.51	0.0583	3.3e-02	0.077		6.01
med.BetaBlockers:followupTime		-0.0073	9.9e-03	0.463		-0.73
med.Epo	0.98	-0.0865	2.8e-02	0.002	**	-8.28
med.Epo:followupTime		0.0082	8.4e-03	0.332		0.82
med.Iron		-0.0285	3.1e-02	0.355		-2.81
med.Iron:followupTime		0.0078	1.1e-02	0.457		0.79
med.ParenteralIron		-0.0518	2.6e-02	0.050	*	-5.04
med.ParenteralIron:followupTime		0.0260	9.0e-03	0.004	**	2.64
med.VitaminD	1.00	-0.1334	3.0e-02	0.000	***	-12.49
med.VitaminD:followupTime		0.0035	8.5e-03	0.685		0.35
numberAKIepisodes	0.80	-0.0155	1.2e-02	0.207		-1.54
numberAKIepisodes:followupTime		0.0016	2.7e-03	0.550		0.16
numberAntihypertensives	0.86	-0.0394	1.7e-02	0.018	*	-3.87
numberAntihypertensives:followupTime		-0.0025	5.2e-03	0.634		-0.25
PO	1.00	-0.1004	1.4e-02	0.000	***	-9.55
PO:followupTime		0.0057	4.1e-03	0.166		0.57
PP	0.72	0.0290	1.2e-02	0.013	*	2.94
PP:followupTime		-0.0014	4.2e-03	0.733		-0.14
PTH	1.00	-0.0761	1.5e-02	0.000	***	-7.33
PTH:followupTime		0.0020	2.9e-03	0.483		0.20
Pu		0.0256	1.6e-02	0.105		2.59
Pu:followupTime		-0.0174	4.8e-03	0.000	***	-1.73
totalCholesterol	0.60	0.0201	1.5e-02	0.174		2.03
totalCholesterol:followupTime		-0.0006	5.0e-03	0.909		-0.06

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.2 Glomerulonephritis

Average effects

The key *average effects* are:

- Medication med.ACE.ARB is associated with higher levels of eGFR; this indicates better kidney function.
- Lower levels of eGFR are associated with taking CCBs medication and vitamin D supplements, and also PO and PTH. This is reasonable since hypertension (in part treated with CCBs) and vitamin D deficiency are associated with poor kidney function. Similarly higher levels of PO and PTH are associated with poor kidney function.
- Higher levels of Hb and total CO2 are associated with higher levels of eGFR. That is better kidney function. This is reasonable since low levels of Hb and total CO2 are associated with poor kidney function.
- Older baseline age is associated with lower eGFR.
- A higher level of Pu is associated with a higher level of eGFR. This result is counterintuitive since increases in Pu indicate worsening kidney function.

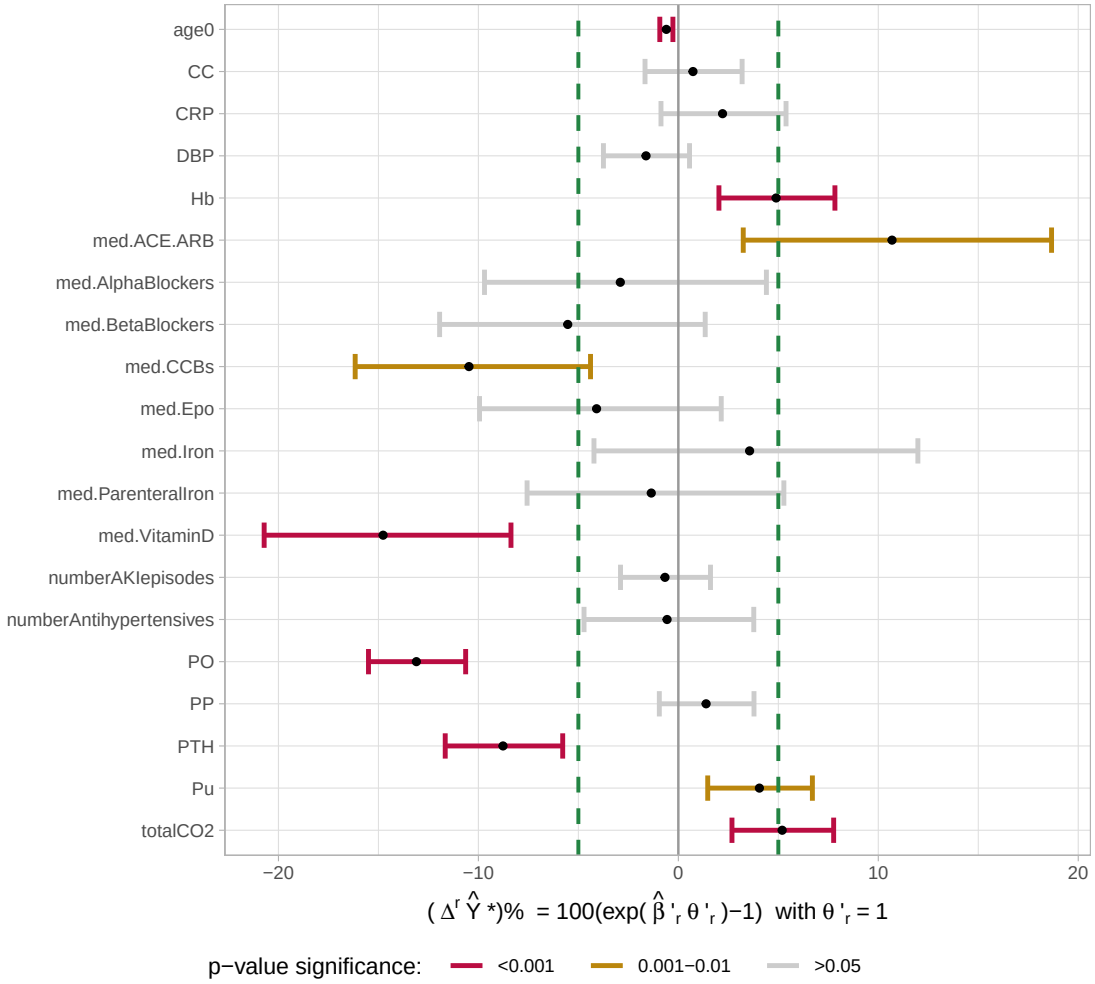

Figure 22: Average effects - relative change in eGFR for standardised model using 95% CIs: glomerulonephritis

Temporal effects

The key *temporal effects* are:

- DBP is associated with a slower decline in eGFR. This is reasonable since higher levels of DBP are associated with better renal function in conjunction with better cardiovascular function.
- Pu is associated with a faster decline in eGFR. This is to be expected since protein in the urine is associated with poorer kidney function.
- PTH is associated with slower progression of kidney disease i.e. shallower slope in eGFR over time. This result is counterintuitive as higher levels of PTH are associated with worsening kidney function.

Although follow-up time is not statistically significant its regression parameter is negative, therefore the level of eGFR is falling off over time. This is to be expected.

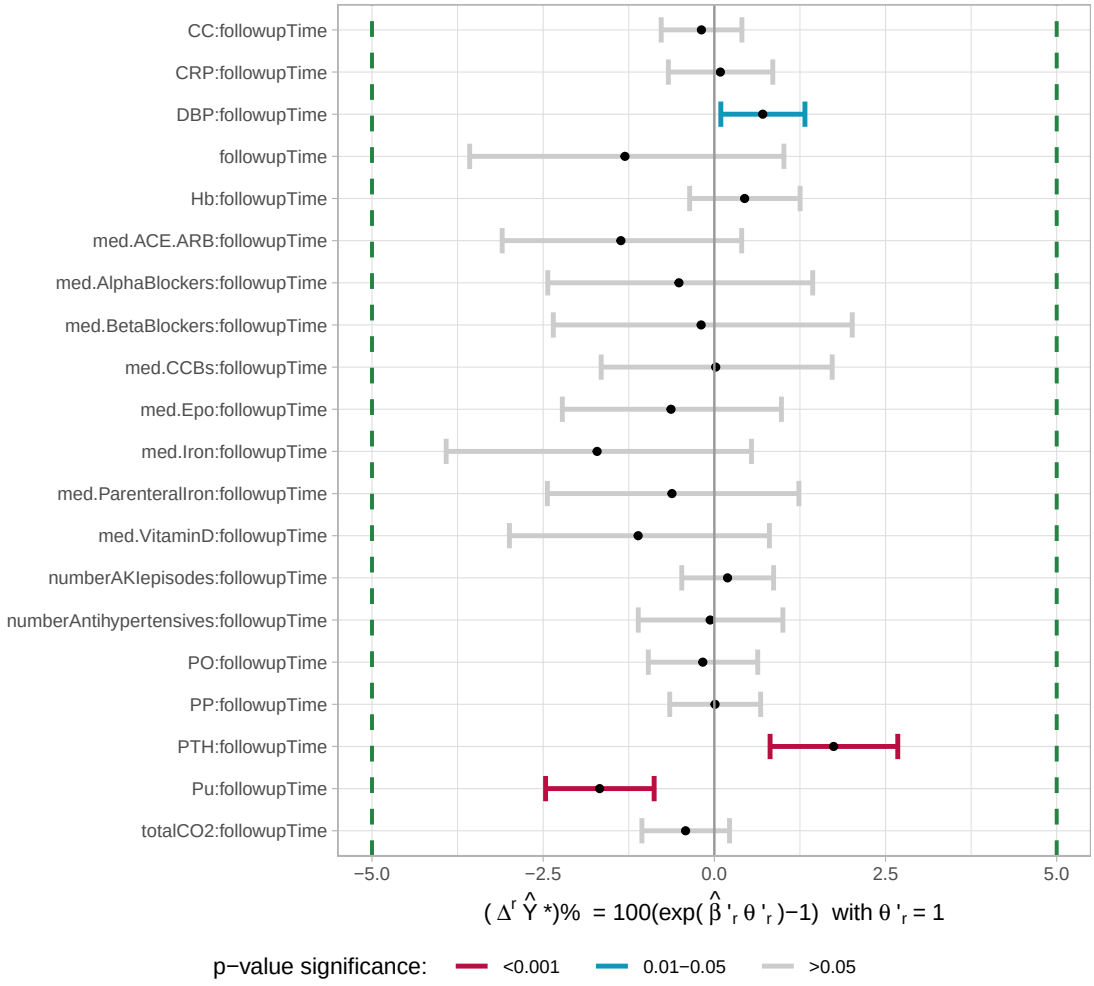

Figure 23: Temporal effects - relative change in eGFR for standardised model using 95% CIs: glomerulonephritis

Parameter values

Table 13: Standardised model summary for disease glomerulonephritis

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	3.8879	1.1e-01	0.000	***	
age0	0.91	-0.0060	1.7e-03	0.000	***	-0.60
CC		0.0073	1.3e-02	0.560		0.73
CC:followupTime		-0.0019	3.1e-03	0.538		-0.19
CRP	0.92	0.0219	1.6e-02	0.169		2.21
CRP:followupTime		0.0009	4.0e-03	0.823		0.09
DBP		-0.0163	1.1e-02	0.152		-1.62
DBP:followupTime		0.0070	3.2e-03	0.026	*	0.71
followupTime	0.96	-0.0131	1.2e-02	0.277		-1.31
Hb	1.00	0.0477	1.4e-02	0.001	**	4.89
Hb:followupTime		0.0044	4.2e-03	0.290		0.44
med.ACE.ARB	0.82	0.1015	3.6e-02	0.005	**	10.69
med.ACE.ARB:followupTime		-0.0137	9.2e-03	0.136		-1.37
med.AlphaBlockers	0.60	-0.0294	3.8e-02	0.434		-2.90
med.AlphaBlockers:followupTime		-0.0052	1.0e-02	0.608		-0.52
med.BetaBlockers	0.76	-0.0569	3.6e-02	0.119		-5.53
med.BetaBlockers:followupTime		-0.0019	1.1e-02	0.865		-0.19
med.CCBs	0.98	-0.1106	3.4e-02	0.001	**	-10.47
med.CCBs:followupTime		0.0002	8.8e-03	0.982		0.02
med.Epo	0.86	-0.0417	3.3e-02	0.202		-4.09
med.Epo:followupTime		-0.0064	8.4e-03	0.447		-0.63
med.Iron		0.0350	4.1e-02	0.388		3.56
med.Iron:followupTime		-0.0173	1.2e-02	0.143		-1.71
med.ParenteralIron		-0.0136	3.4e-02	0.687		-1.35
med.ParenteralIron:followupTime		-0.0062	9.6e-03	0.518		-0.62
med.VitaminD	1.00	-0.1598	3.8e-02	0.000	***	-14.77
med.VitaminD:followupTime		-0.0112	1.0e-02	0.262		-1.11
numberAKIepisodes	0.53	-0.0067	1.2e-02	0.569		-0.67
numberAKIepisodes:followupTime		0.0019	3.5e-03	0.580		0.19
numberAntihypertensives		-0.0056	2.2e-02	0.799		-0.56
numberAntihypertensives:followupTime		-0.0006	5.5e-03	0.912		-0.06
PO	1.00	-0.1404	1.5e-02	0.000	***	-13.10
PO:followupTime		-0.0017	4.2e-03	0.685		-0.17
PP		0.0138	1.2e-02	0.256		1.39
PP:followupTime		0.0001	3.4e-03	0.977		0.01
PTH	0.98	-0.0918	1.7e-02	0.000	***	-8.77
PTH:followupTime		0.0173	4.8e-03	0.000	***	1.74
Pu		0.0397	1.3e-02	0.002	**	4.05
Pu:followupTime		-0.0169	4.2e-03	0.000	***	-1.68
totalCO2	1.00	0.0506	1.3e-02	0.000	***	5.19
totalCO2:followupTime		-0.0042	3.3e-03	0.207		-0.42

Table 13: Standardised model summary for disease glomerulonephritis *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.3 Hypertensive kidney disease

Average effects

The key *average effects* are:

- Taking higher numbers of antihypertensive drugs is associated with lower eGFR levels. This is reasonable because poor kidney function is known to be associated with hypertension.
- Having more than one type of cardiovascular (CV) disease is associated with lower levels of eGFR. This is consistent given poorer kidney function, which is associated with an increase in risk of CV disease (and vice versa).
- Lower levels of eGFR are associated with taking vitamin D supplement, and also higher levels of PO and PTH. Vitamin D deficiency is associated with poor kidney function. Similarly higher levels of PO and PTH are associated with poor kidney function.
- Higher levels of Hb and total CO2 are associated with higher levels of eGFR. Note that low levels of both these biochemicals are associated with poor kidney function.
- Patients who are older at baseline have poorer kidney function.

Considering the effects with a significance of 0.01-0.05 we find that drinking more than 1 unit of alcohol per week and a previous cancer are both associated with lower eGFR levels. Note that the cancer category is very general as it includes all types of cancer and is not confined to renal cancer.

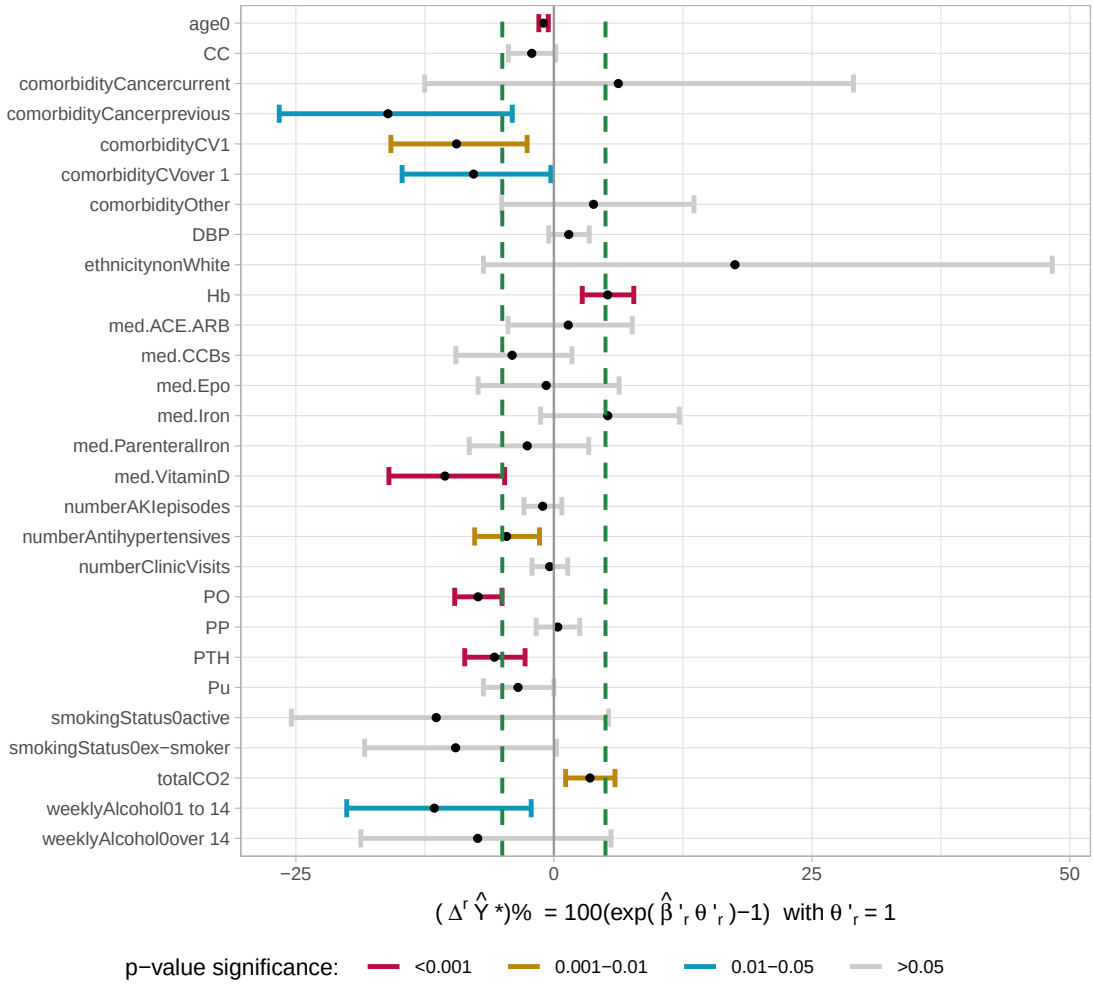

Figure 24: Average effects - relative change in eGFR for standardised model using 95% CIs: HKD

Temporal effects

The key *temporal effects* are:

- A high number of clinic visits is associated with a more rapid decline in kidney function. This would imply that patients with poorer health visit the clinic more frequently.
- Pu is associated with a more rapid decline in kidney function. The presence of protein in the urine is associated with poor kidney function.

In addition, follow-up time is negative and significant at the 0.01-0.05 level. This means that the level of eGFR is dropping off over time, which is to be expected. The significance of follow-up time may indicate that there are key risk factors which are not included in our model.

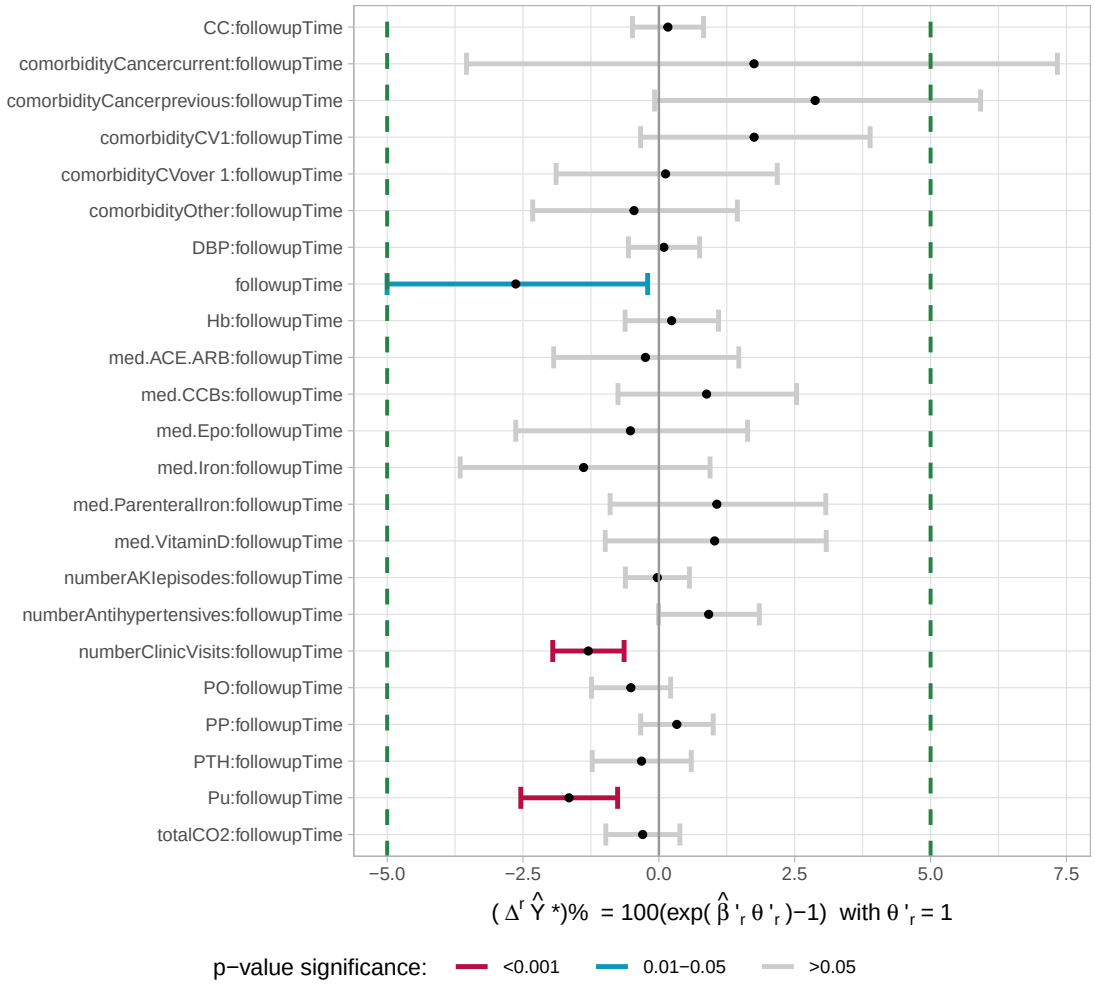

Figure 25: Temporal effects - relative change in eGFR for standardised model using 95% CIs: HKD

Parameter values

Table 14: Standardised model summary for disease HKD

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	4.2133	1.8e-01	0.000	***	
age0	0.97	-0.0100	2.4e-03	0.000	***	-1.00
CC	0.57	-0.0216	1.2e-02	0.079		-2.14
CC:followupTime		0.0017	3.4e-03	0.628		0.17
comorbidityCancercurrent	0.77	0.0606	1.0e-01	0.553		6.24
comorbidityCancercurrent:followupTime		0.0174	2.8e-02	0.536		1.75
comorbidityCancerprevious	0.77	-0.1753	7.0e-02	0.013	*	-16.08
comorbidityCancerprevious:followupTime		0.0284	1.5e-02	0.064		2.88
comorbidityCV1	0.70	-0.0991	3.8e-02	0.010	*	-9.43
comorbidityCV1:followupTime		0.0174	1.1e-02	0.111		1.75
comorbidityCVover 1	0.70	-0.0810	4.1e-02	0.048	*	-7.78
comorbidityCVover 1:followupTime		0.0012	1.1e-02	0.908		0.12
comorbidityOther	0.61	0.0377	4.7e-02	0.423		3.84
comorbidityOther:followupTime		-0.0046	9.9e-03	0.645		-0.46
DBP		0.0143	1.0e-02	0.158		1.44
DBP:followupTime		0.0009	3.4e-03	0.788		0.09
ethnicitynonWhite	0.55	0.1617	1.2e-01	0.185		17.55
followupTime	0.67	-0.0267	1.3e-02	0.039	*	-2.63
Hb	1.00	0.0508	1.2e-02	0.000	***	5.22
Hb:followupTime		0.0023	4.5e-03	0.603		0.23
med.ACE.ARB		0.0138	3.1e-02	0.657		1.39
med.ACE.ARB:followupTime		-0.0025	9.0e-03	0.783		-0.25
med.CCBs	0.64	-0.0413	3.1e-02	0.180		-4.04
med.CCBs:followupTime		0.0087	8.6e-03	0.307		0.88
med.Epo		-0.0075	3.6e-02	0.836		-0.74
med.Epo:followupTime		-0.0052	1.1e-02	0.642		-0.52
med.Iron		0.0509	3.3e-02	0.129		5.22
med.Iron:followupTime		-0.0139	1.2e-02	0.255		-1.38
med.ParenteralIron		-0.0261	3.1e-02	0.402		-2.58
med.ParenteralIron:followupTime		0.0106	1.0e-02	0.303		1.07
med.VitaminD	0.75	-0.1115	3.3e-02	0.001	**	-10.55
med.VitaminD:followupTime		0.0102	1.1e-02	0.334		1.03
numberAKIepisodes	0.69	-0.0109	9.7e-03	0.263		-1.08
numberAKIepisodes:followupTime		-0.0003	3.1e-03	0.925		-0.03
numberAntihypertensives	0.64	-0.0469	1.7e-02	0.007	**	-4.59
numberAntihypertensives:followupTime		0.0091	4.8e-03	0.059		0.92
numberClinicVisits	0.70	-0.0041	9.1e-03	0.654		-0.41
numberClinicVisits:followupTime		-0.0131	3.5e-03	0.000	***	-1.30
PO	1.00	-0.0763	1.3e-02	0.000	***	-7.35
PO:followupTime		-0.0052	3.8e-03	0.180		-0.51
PP		0.0037	1.1e-02	0.736		0.37

Table 14: Standardised model summary for disease HKD *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
PP:followupTime		0.0033	3.5e-03	0.345		0.33
PTH	1.00	-0.0593	1.6e-02	0.000	***	-5.76
PTH:followupTime		-0.0032	4.8e-03	0.505		-0.32
Pu	1.00	-0.0353	1.9e-02	0.058		-3.47
Pu:followupTime		-0.0167	4.8e-03	0.000	***	-1.65
smokingStatus0active	0.94	-0.1210	9.0e-02	0.182		-11.40
smokingStatus0ex-smoker	0.94	-0.1001	5.4e-02	0.063		-9.52
totalCO2	0.85	0.0344	1.2e-02	0.005	**	3.50
totalCO2:followupTime		-0.0030	3.6e-03	0.406		-0.30
weeklyAlcohol01 to 14	0.65	-0.1232	5.3e-02	0.021	*	-11.59
weeklyAlcohol0over 14	0.65	-0.0767	6.8e-02	0.263		-7.38

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.4 Other

Average effects

The key *average effects* are:

- Lower levels of eGFR are associated with taking a vitamin D supplement and also higher levels of PO and PTH. This is reasonable since vitamin D deficiency and high levels of PO and PTH are associated with poor kidney function.
- Higher levels of Hb and total CO2 are associated with higher levels of eGFR, meaning that the kidney function is better. Poor kidney function is associated with low levels of both of these biochemicals.
- Patients who are older at baseline have poorer kidney function.

We also note, at the 0.01-0.05 significance level, that a higher number of clinic visits is associated with higher levels of eGFR. This may indicate that the decline in renal function for these patients is being better controlled by more frequent monitoring of their condition. Furthermore we observe that lower levels of CC are associated with lower levels of eGFR. This is medically plausible.

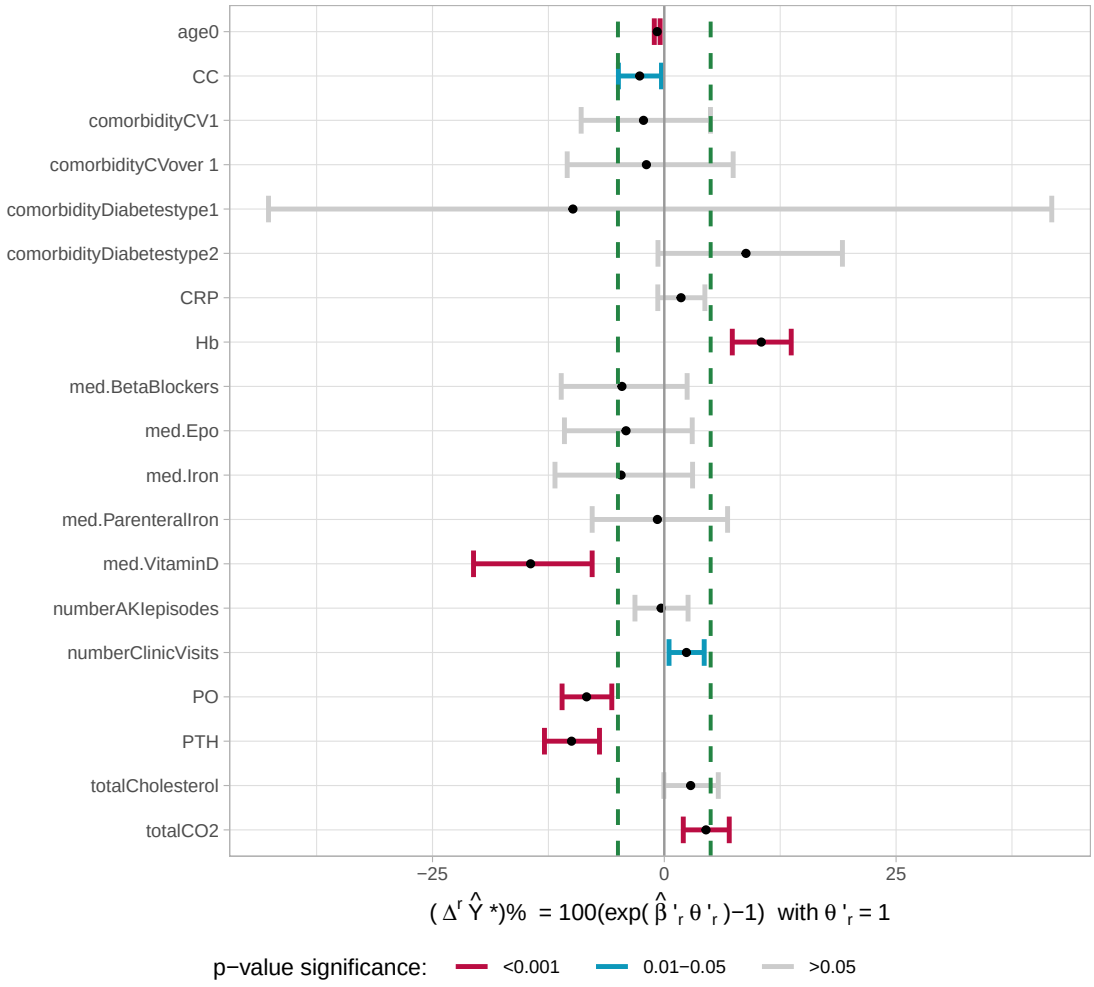

Figure 26: Average effects - relative change in eGFR for standardised model using 95% CIs: other

Temporal effects

The key *temporal effects* are:

- PO is associated with a more rapid decline in kidney function.
- PTH is associated with slower progression of kidney disease. This result is counterintuitive as higher levels of PTH tend to occur with worsening kidney function.

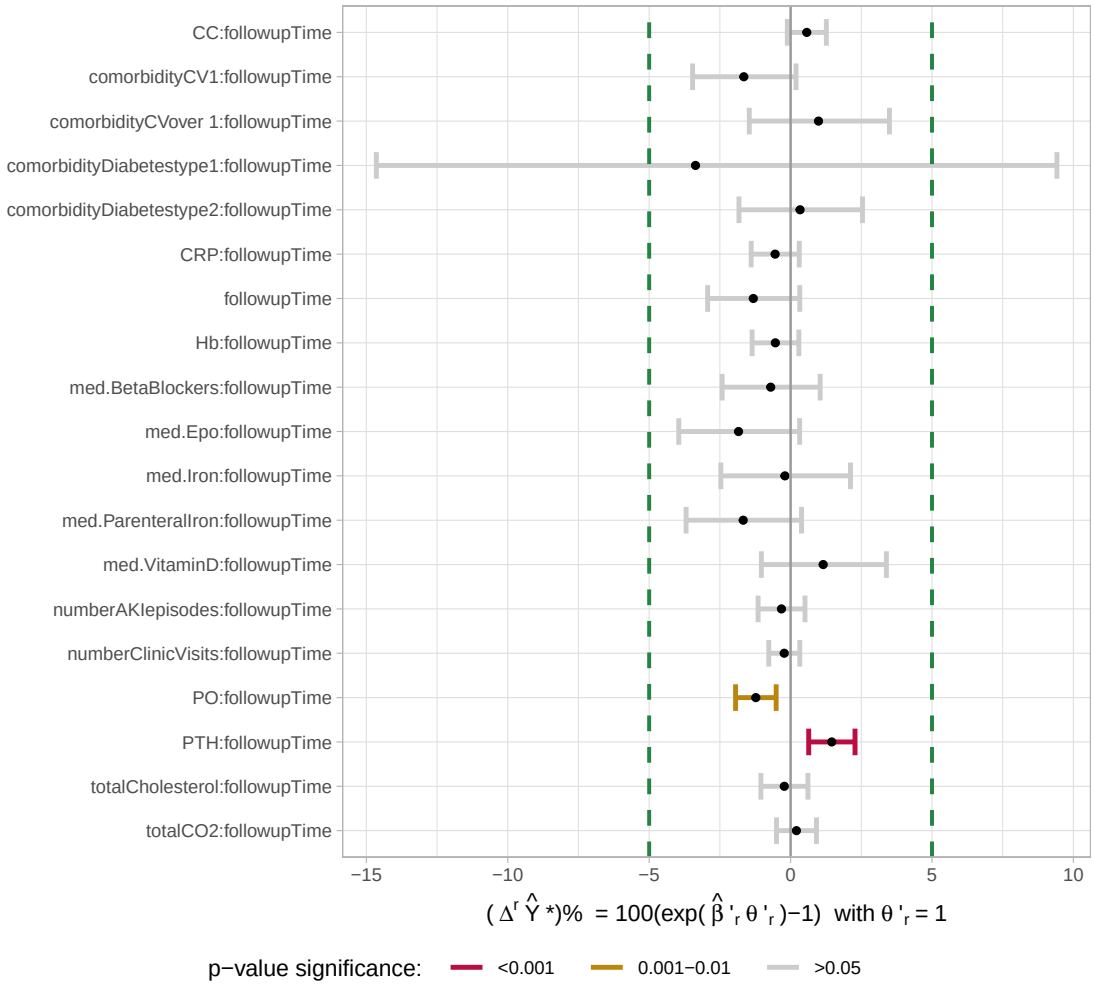

Figure 27: Temporal effects - relative change in eGFR for standardised model using 95% CIs: disease other

Parameter values

Table 15: Standardised model summary for disease other

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	4.0023	1.0e-01	0.000	***	
age0	1.00	-0.0077	1.7e-03	0.000	***	-0.76
CC		-0.0269	1.2e-02	0.030	*	-2.66
CC:followupTime		0.0057	3.6e-03	0.110		0.57
comorbidityCV1	0.65	-0.0227	3.7e-02	0.540		-2.24
comorbidityCV1:followupTime		-0.0167	9.7e-03	0.085		-1.65
comorbidityCVover 1	0.65	-0.0195	4.7e-02	0.681		-1.93
comorbidityCVover 1:followupTime		0.0098	1.3e-02	0.442		0.99
comorbidityDiabetestype1	0.68	-0.1038	2.4e-01	0.660		-9.86
comorbidityDiabetestype1:followupTime		-0.0342	6.5e-02	0.596		-3.36
comorbidityDiabetestype2	0.91	0.0844	4.7e-02	0.076		8.81
comorbidityDiabetestype2:followupTime		0.0033	1.1e-02	0.767		0.34
CRP	0.60	0.0179	1.3e-02	0.168		1.81
CRP:followupTime		-0.0055	4.4e-03	0.218		-0.55
followupTime	0.99	-0.0133	8.6e-03	0.122		-1.32
Hb	1.00	0.0995	1.5e-02	0.000	***	10.47
Hb:followupTime		-0.0054	4.3e-03	0.212		-0.54
med.BetaBlockers	0.52	-0.0468	3.7e-02	0.206		-4.57
med.BetaBlockers:followupTime		-0.0071	9.1e-03	0.437		-0.70
med.Epo	0.89	-0.0422	3.7e-02	0.259		-4.13
med.Epo:followupTime		-0.0186	1.1e-02	0.101		-1.84
med.Iron		-0.0477	4.0e-02	0.238		-4.66
med.Iron:followupTime		-0.0020	1.2e-02	0.867		-0.20
med.ParenteralIron		-0.0075	3.8e-02	0.845		-0.75
med.ParenteralIron:followupTime		-0.0169	1.1e-02	0.118		-1.68
med.VitaminD	0.96	-0.1557	3.9e-02	0.000	***	-14.42
med.VitaminD:followupTime		0.0115	1.1e-02	0.313		1.15
numberAKIepisodes	0.65	-0.0034	1.5e-02	0.819		-0.34
numberAKIepisodes:followupTime		-0.0032	4.3e-03	0.454		-0.32
numberClinicVisits	0.86	0.0236	9.6e-03	0.014	*	2.39
numberClinicVisits:followupTime		-0.0022	2.9e-03	0.432		-0.22
PO	1.00	-0.0875	1.5e-02	0.000	***	-8.38
PO:followupTime		-0.0124	3.8e-03	0.001	**	-1.23
PTH	0.98	-0.1054	1.7e-02	0.000	***	-10.01
PTH:followupTime		0.0145	4.2e-03	0.001	**	1.46
totalCholesterol	0.52	0.0281	1.5e-02	0.059		2.84
totalCholesterol:followupTime		-0.0022	4.3e-03	0.608		-0.22
totalCO2	1.00	0.0440	1.2e-02	0.000	***	4.50
totalCO2:followupTime		0.0021	3.7e-03	0.574		0.21

Table 15: Standardised model summary for disease other *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.5 PKD

Average effects

The key *average effects* are:

- Older age is associated with lower levels of eGFR.
- Lower levels of eGFR are associated with higher levels of PTH. Note that higher PTH is associated with poorer kidney function.

At a statistical significance of 0.01-0.05 we also note the following results:

- Lower levels of eGFR are associated with patients receiving parenteral iron and also higher levels of PO and Pu. Kidney disease can result in anaemia so patients receiving parenteral iron would be expected to have lower levels of eGFR. Higher levels of PO and Pu are associated with poorer kidney function.
- Total CO2 is associated with higher levels of eGFR, note that poor kidney function can cause low levels of total CO2.
- PP is associated with higher levels of eGFR. It is expected that PP will increase with both age and worsening renal function therefore this result is counterintuitive.

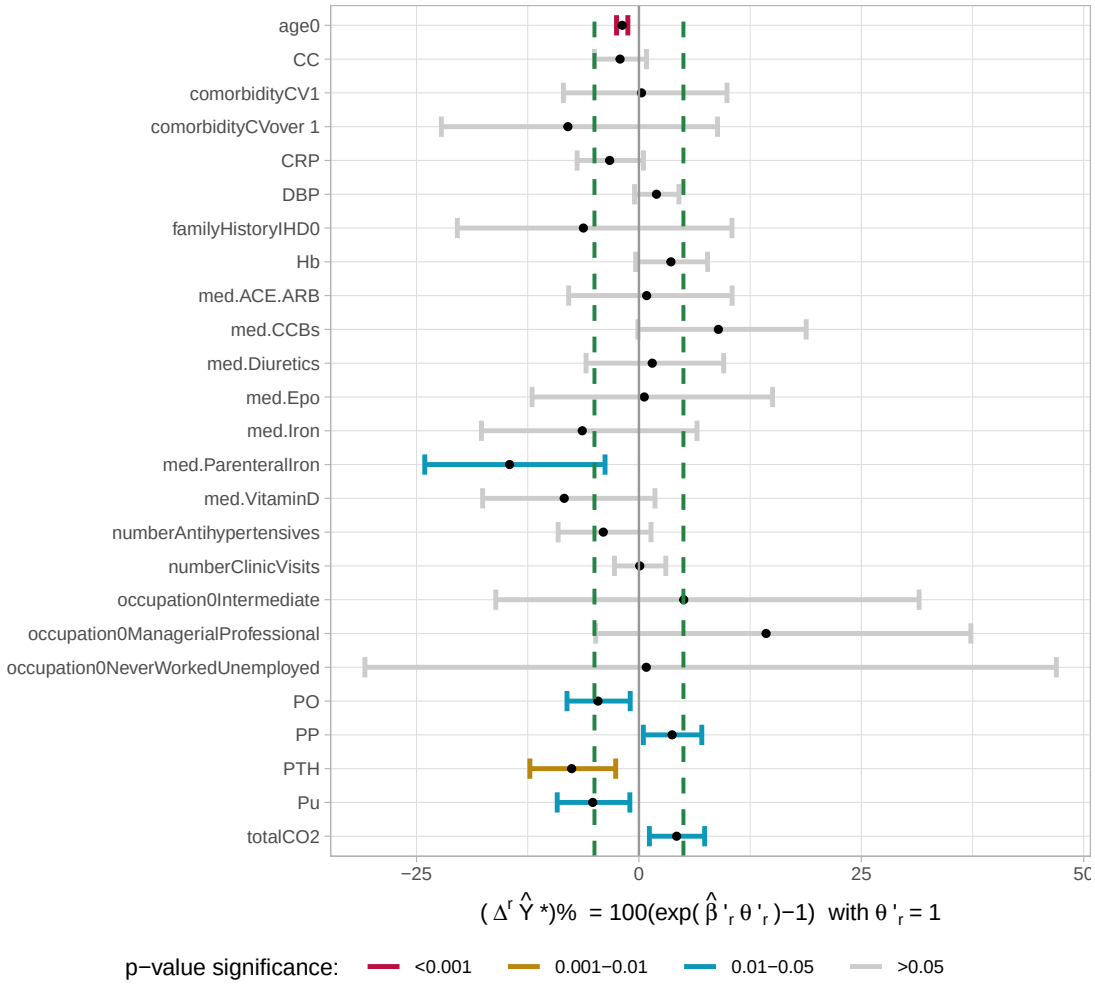

Figure 28: Average effects - relative change in eGFR for standardised model using 95% CIs: PKD

Temporal effects

The key *temporal effects* are:

- Follow-up time is negative which indicates that the level of eGFR is dropping off over time. Follow-up time is very strongly associated with lower levels of eGFR. This may suggest that the model is missing at least one risk factor. In PKD patients the continued growth of cysts in the kidneys progressively impairs their function. Our model does not include variables for the size, growth rate, and/or number of cysts in the kidneys. If it included such variables it is possible that follow-up time would be either less significant or not significant.
- Hb, although only significant at the 0.01-0.05 level, is associated with slower progression of CKD. This result is consistent with PKD patients tending to maintain good levels of Hb until very late in their disease progression.

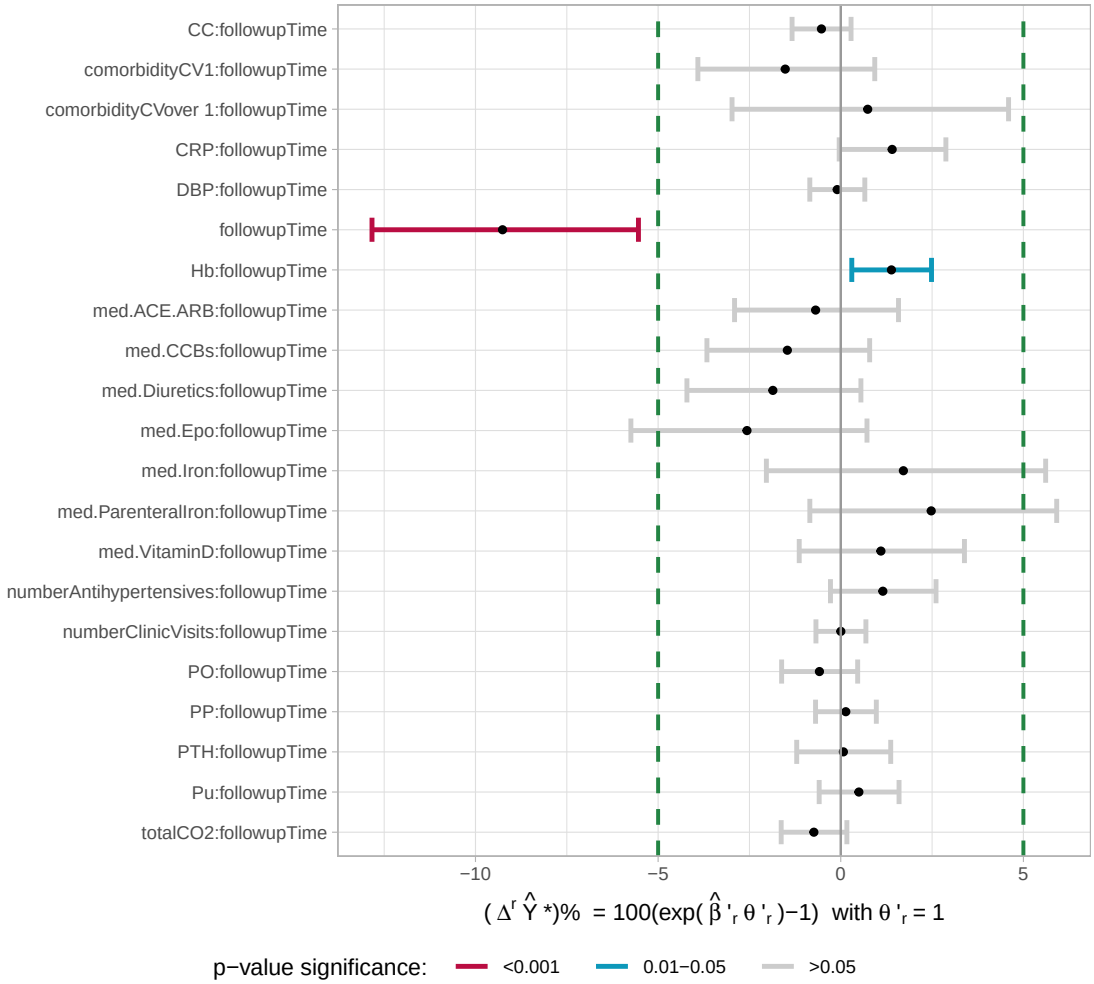

Figure 29: Temporal effects - relative change in eGFR for standardised model using 95% CIs: PKD

Parameter values

Table 16: Standardised model summary for disease polycystic kidney disease

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	4.3665	2.1e-01	0.000	***	
age0	0.98	-0.0190	3.5e-03	0.000	***	-1.88
CC	0.90	-0.0214	1.6e-02	0.186		-2.12
CC:followupTime		-0.0053	4.4e-03	0.228		-0.53
comorbidityCV1	0.70	0.0029	4.9e-02	0.954		0.29
comorbidityCV1:followupTime		-0.0153	1.3e-02	0.250		-1.52
comorbidityCVover 1	0.70	-0.0832	9.1e-02	0.360		-7.99
comorbidityCVover 1:followupTime		0.0073	2.0e-02	0.718		0.74
CRP	0.60	-0.0335	2.1e-02	0.109		-3.29
CRP:followupTime		0.0140	7.8e-03	0.074		1.41
DBP	0.59	0.0195	1.3e-02	0.141		1.97
DBP:followupTime		-0.0010	4.1e-03	0.811		-0.10
familyHistoryIHD0	0.58	-0.0645	8.8e-02	0.465		-6.24
followupTime	1.00	-0.0972	2.2e-02	0.000	***	-9.26
Hb	0.90	0.0353	2.1e-02	0.095		3.60
Hb:followupTime		0.0138	5.8e-03	0.019	*	1.39
med.ACE.ARB		0.0087	4.9e-02	0.860		0.87
med.ACE.ARB:followupTime		-0.0069	1.2e-02	0.572		-0.69
med.CCBs	0.64	0.0856	4.7e-02	0.069		8.93
med.CCBs:followupTime		-0.0147	1.2e-02	0.229		-1.46
med.Diuretics	0.56	0.0149	4.1e-02	0.718		1.50
med.Diuretics:followupTime		-0.0188	1.3e-02	0.154		-1.86
med.Epo	0.54	0.0060	7.2e-02	0.934		0.60
med.Epo:followupTime		-0.0260	1.8e-02	0.148		-2.57
med.Iron		-0.0658	7.0e-02	0.346		-6.37
med.Iron:followupTime		0.0170	2.0e-02	0.403		1.71
med.ParenteralIron	0.69	-0.1571	6.4e-02	0.015	*	-14.54
med.ParenteralIron:followupTime		0.0245	1.8e-02	0.171		2.48
med.VitaminD	0.81	-0.0877	5.7e-02	0.125		-8.40
med.VitaminD:followupTime		0.0109	1.2e-02	0.367		1.10
numberAntihypertensives		-0.0409	2.9e-02	0.165		-4.00
numberAntihypertensives:followupTime		0.0115	7.7e-03	0.139		1.15
numberClinicVisits	0.64	0.0009	1.6e-02	0.952		0.09
numberClinicVisits:followupTime		0.0000	3.7e-03	0.996		0.00
occupation0ManagerialProfessional	0.59	0.1336	9.8e-02	0.178		14.29
occupation0Intermediate	0.59	0.0491	1.2e-01	0.684		5.04
occupation0NeverWorkedUnemployed	0.59	0.0082	2.0e-01	0.968		0.83
PO	0.98	-0.0471	2.0e-02	0.020	*	-4.60
PO:followupTime		-0.0059	5.7e-03	0.302		-0.58
PP	0.92	0.0366	1.7e-02	0.033	*	3.73
PP:followupTime		0.0014	4.5e-03	0.759		0.14

Table 16: Standardised model summary for disease polycystic kidney disease *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
PTH	1.00	-0.0786	2.8e-02	0.006	**	-7.56
PTH:followupTime		0.0007	6.9e-03	0.917		0.07
Pu	0.75	-0.0533	2.3e-02	0.023	*	-5.19
Pu:followupTime		0.0049	5.9e-03	0.401		0.50
totalCO2	0.79	0.0415	1.6e-02	0.011	*	4.23
totalCO2:followupTime		-0.0074	4.9e-03	0.132		-0.74

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.6 Pyelonephritis

Average effects

The key *average effects* are:

- Lower levels of eGFR are associated with patients taking vitamin D supplements, larger numbers of antihypertensives, and also PO. Kidney disease can cause vitamin D deficiency and hypertension so patients taking Vitamin D and a higher number of antihypertensives are expected to have lower levels of eGFR. Patients with higher levels PO will have poorer kidney function and therefore lower levels of eGFR.
- Total CO2 is associated with higher levels of eGFR. Note that poor kidney function can cause lower levels of total CO2.
- An older age at baseline is associated with a lower level of eGFR.

Factors at the 0.01-0.05 significance level are as follows:

- Hb is associated with higher levels of eGFR. Note that poor kidney function may cause low levels of Hb.
- Drinking more than 14 units of alcohol per week is associated with higher levels of kidney function. We consider this to be an anomaly due to the wide confidence intervals and the fact that alcohol consumption is not expected to be associated with higher levels of eGFR.

We also note, although it is not significant, that this model selected the *sexfemale* variable. Being female is weakly associated with lower levels of eGFR. This is reasonable because urinary tract infections are more common in females. These infections can lead to kidney damage and in particular pyelonephritis.

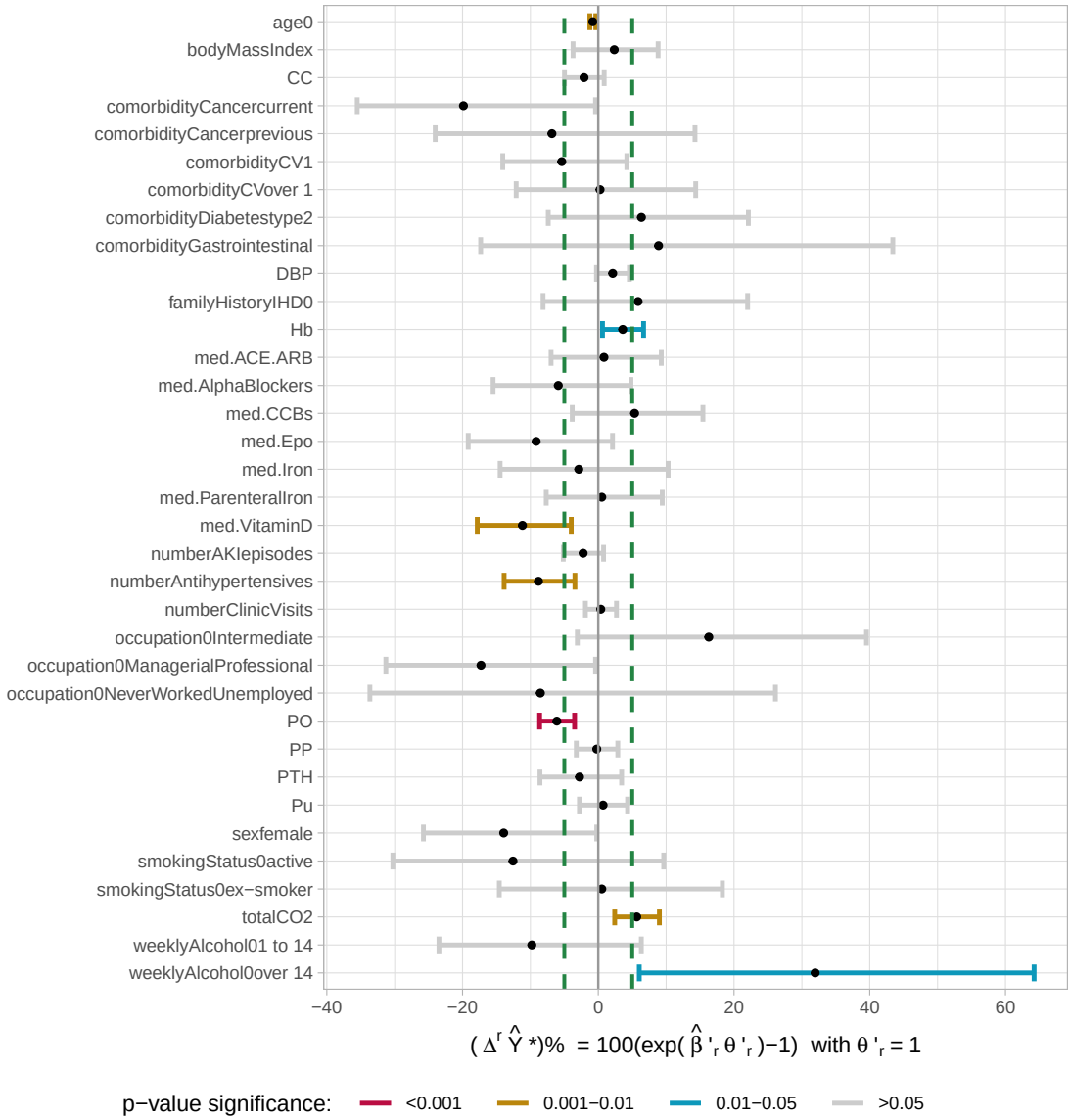

Figure 30: Average effects - relative change in eGFR for standardised model using 95% CIs: pyelonephritis

Temporal effects

The key *temporal effects* are:

- Hb is associated with a slower progression of kidney disease.
- Pu and total CO2 are associated with a more rapid progression of kidney disease.

Note that although not statistically significant the follow-up time regression parameter is negative. This indicates that eGFR falls over time.

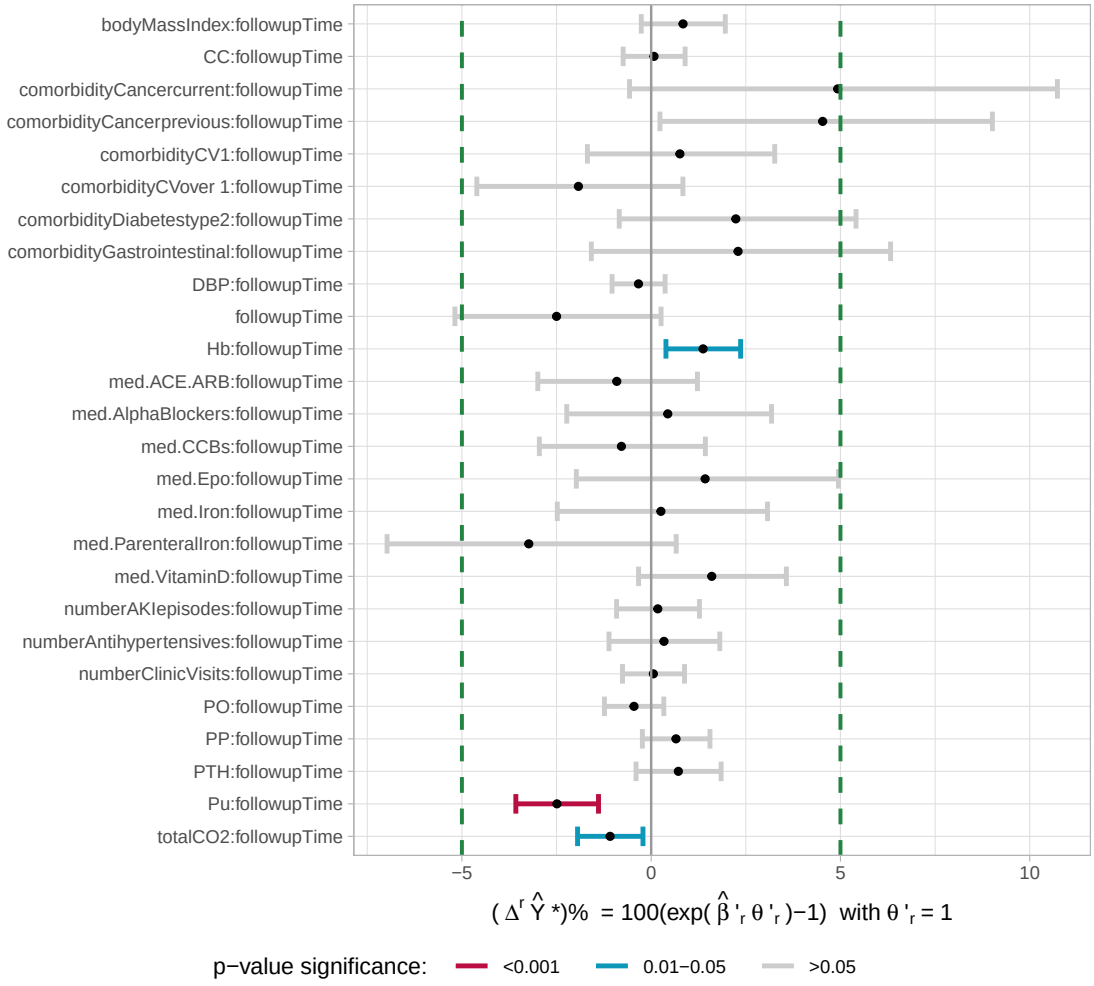

Figure 31: Temporal effects - relative change in eGFR for standardised model using 95% CIs: pyelonephritis

Parameter values

Table 17: Standardised model summary for disease pyelonephritis

parameter	prop[a]	$\hat{\beta}_r'$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	3.9291	1.8e-01	0.000	***	
age0	0.86	-0.0081	2.4e-03	0.001	**	-0.81
bodyMassIndex	0.74	0.0233	3.3e-02	0.486		2.36
bodyMassIndex:followupTime		0.0084	6.0e-03	0.164		0.84
CC		-0.0212	1.6e-02	0.194		-2.10
CC:followupTime		0.0007	4.5e-03	0.867		0.07
comorbidityCancercurrent	0.70	-0.2215	1.2e-01	0.063		-19.87
comorbidityCancercurrent:followupTime		0.0481	2.9e-02	0.103		4.93
comorbidityCancerprevious	0.71	-0.0708	1.1e-01	0.526		-6.83
comorbidityCancerprevious:followupTime		0.0443	2.3e-02	0.054		4.53
comorbidityCV1	0.53	-0.0553	5.3e-02	0.295		-5.38
comorbidityCV1:followupTime		0.0076	1.3e-02	0.573		0.76
comorbidityCVover 1	0.53	0.0026	7.2e-02	0.971		0.26
comorbidityCVover 1:followupTime		-0.0194	1.5e-02	0.202		-1.92
comorbidityDiabetestype2	0.58	0.0616	7.5e-02	0.415		6.35
comorbidityDiabetestype2:followupTime		0.0221	1.7e-02	0.187		2.24
comorbidityGastrointestinal	0.84	0.0851	1.5e-01	0.570		8.88
comorbidityGastrointestinal:followupTime		0.0227	2.1e-02	0.283		2.30
DBP	0.59	0.0211	1.3e-02	0.105		2.13
DBP:followupTime		-0.0034	3.8e-03	0.383		-0.33
familyHistoryIHD0	0.55	0.0570	7.7e-02	0.460		5.87
followupTime	0.51	-0.0253	1.5e-02	0.098		-2.50
Hb	0.98	0.0354	1.6e-02	0.027	*	3.60
Hb:followupTime		0.0136	5.3e-03	0.011	*	1.37
med.ACE.ARB		0.0083	4.4e-02	0.850		0.84
med.ACE.ARB:followupTime		-0.0091	1.2e-02	0.433		-0.91
med.AlphaBlockers	0.66	-0.0607	5.9e-02	0.303		-5.89
med.AlphaBlockers:followupTime		0.0044	1.5e-02	0.767		0.44
med.CCBs	0.61	0.0521	5.0e-02	0.296		5.35
med.CCBs:followupTime		-0.0079	1.2e-02	0.514		-0.79
med.Epo		-0.0960	6.4e-02	0.133		-9.15
med.Epo:followupTime		0.0142	1.9e-02	0.448		1.43
med.Iron		-0.0292	6.9e-02	0.675		-2.87
med.Iron:followupTime		0.0026	1.5e-02	0.865		0.26
med.ParenteralIron		0.0051	4.6e-02	0.912		0.51
med.ParenteralIron:followupTime		-0.0329	2.2e-02	0.128		-3.23
med.VitaminD	0.99	-0.1185	4.3e-02	0.006	**	-11.17
med.VitaminD:followupTime		0.0159	1.0e-02	0.132		1.60
numberAKIepisodes	0.60	-0.0226	1.7e-02	0.174		-2.23
numberAKIepisodes:followupTime		0.0017	6.0e-03	0.770		0.17
numberAntihypertensives	0.85	-0.0922	3.1e-02	0.003	**	-8.81

Table 17: Standardised model summary for disease pyelonephritis *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
numberAntihypertensives:followupTime		0.0034	8.0e-03	0.672		0.34
numberClinicVisits	0.52	0.0037	1.2e-02	0.768		0.37
numberClinicVisits:followupTime		0.0006	4.5e-03	0.897		0.06
occupation0ManagerialProfessional	0.72	-0.1897	1.0e-01	0.062		-17.28
occupation0Intermediate	0.72	0.1509	9.9e-02	0.130		16.28
occupation0NeverWorkedUnemployed	0.72	-0.0892	1.7e-01	0.609		-8.53
PO	1.00	-0.0630	1.5e-02	0.000	***	-6.10
PO:followupTime		-0.0046	4.3e-03	0.291		-0.45
PP		-0.0023	1.7e-02	0.893		-0.23
PP:followupTime		0.0065	4.8e-03	0.178		0.66
PTH	0.74	-0.0280	3.4e-02	0.407		-2.76
PTH:followupTime		0.0072	6.1e-03	0.241		0.72
Pu	0.93	0.0070	1.9e-02	0.715		0.71
Pu:followupTime		-0.0252	6.1e-03	0.000	***	-2.49
sexfemale	0.81	-0.1501	8.0e-02	0.064		-13.94
smokingStatus0active	0.93	-0.1343	1.2e-01	0.276		-12.56
smokingStatus0ex-smoker	0.93	0.0052	8.8e-02	0.954		0.52
totalCO2	0.67	0.0551	1.7e-02	0.001	**	5.66
totalCO2:followupTime		-0.0109	4.8e-03	0.023	*	-1.08
weeklyAlcohol01 to 14	0.82	-0.1031	8.9e-02	0.251		-9.79
weeklyAlcohol0over 14	0.82	0.2773	1.2e-01	0.021	*	31.96

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.7 Renovascular

Average effects

The key *average effects* are:

- Hb is associated with higher levels of eGFR. Note that poor kidney function can result in lower levels of Hb.
- EPO treatment is all associated with lower levels of eGFR. Patients requiring this treatment will generally have poorer kidney function.
- Higher levels of PO and PTH are associated with lower levels of eGFR and poorer kidney function.
- PP is associated with higher levels of eGFR. It is expected that PP will increase with both age and worsening renal function therefore this result is counterintuitive.

Although not as significant as the factors listed above, having more then one cardiovascular (CV) disease, taking diuretic medications, having an older baseline age and taking vitamin D supplements are all associated with lower levels of eGFR. Similarly taking parenteral iron is associated with better kidney function.

Note that the very wide confidence interval on baseline occupation *NeverWorkedUnemplyed* is due to there being very few realisations of this factor level. Out of about 560 realisations this variable has about 5; the exact numbers are documented in Appendix A.6.

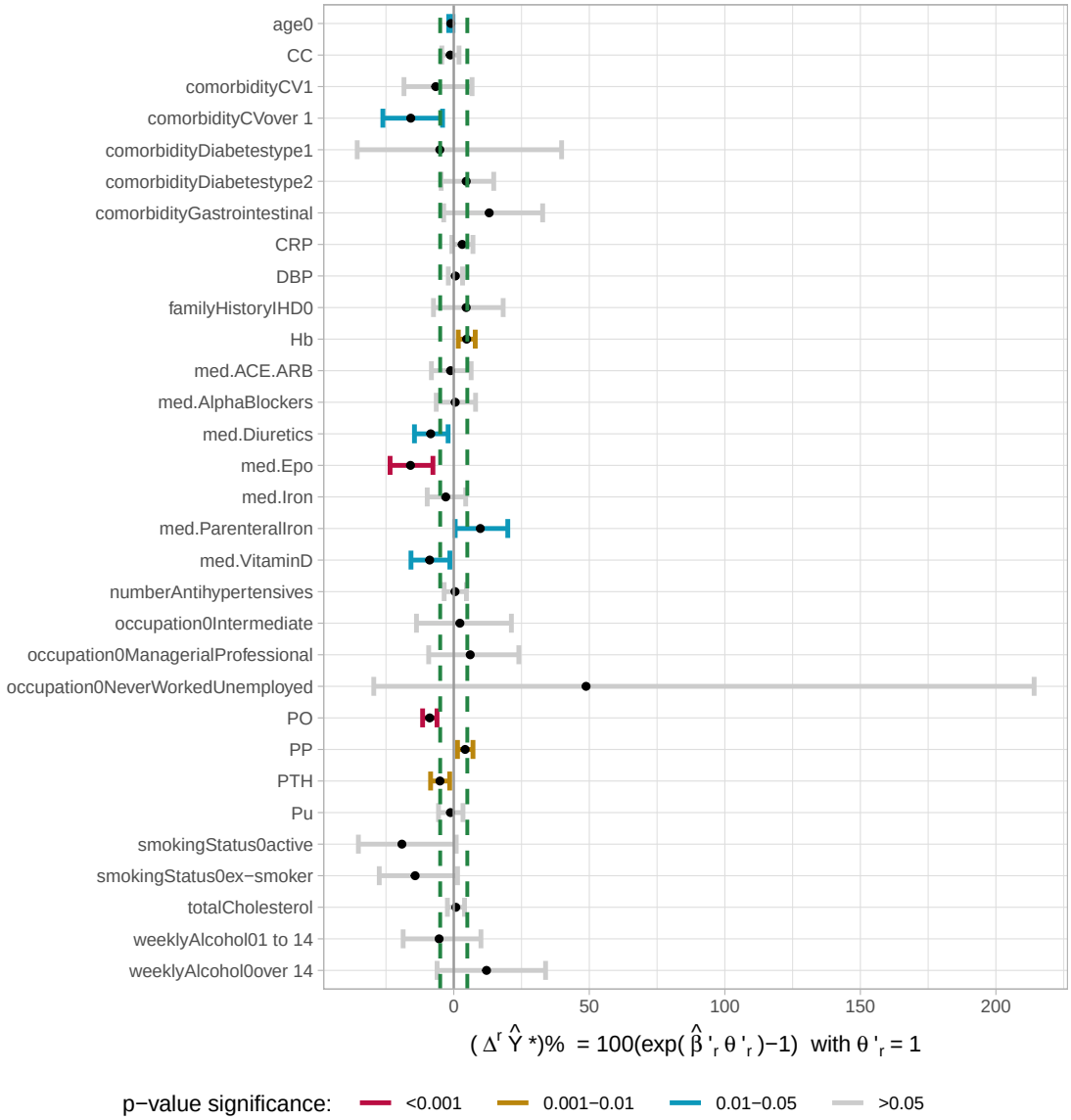

Figure 32: Average effects - relative change in eGFR for standardised model using 95% CIs: renovascular

Temporal effects

The key *temporal effects* are:

- Alpha and/or beta blockers are associated with a less rapid decline in kidney function.
- Total cholesterol is associated with a more rapid decline in kidney function.

DBP, with significance level of 0.01-0.05, is relatively weakly associated with a less rapid decline in kidney function. This is unexpected because as PP rises (e.g. with age) the clinicians expect DBP to fall.

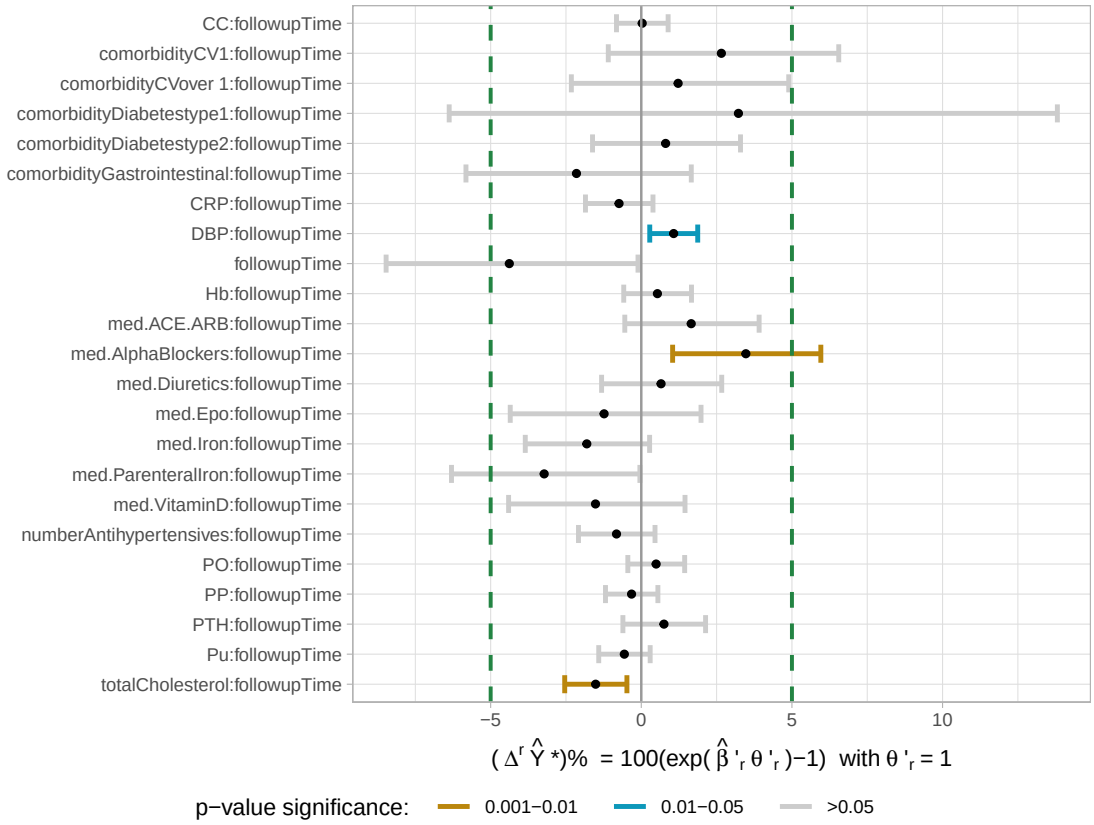

Figure 33: Temporal effects - relative change in eGFR for standardised model using 95% CIs: renovascular

Parameter values

Table 18: Standardised model summary for disease renovascular

parameter	prop[a]	$\hat{\beta}_r'$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	4.3841	3.3e-01	0.000	***	
age0	0.73	-0.0109	4.2e-03	0.011	*	-1.08
CC		-0.0129	1.7e-02	0.453		-1.28
CC:followupTime		0.0003	4.6e-03	0.947		0.03
comorbidityCV1	0.69	-0.0688	7.2e-02	0.339		-6.64
comorbidityCV1:followupTime		0.0262	2.0e-02	0.189		2.66
comorbidityCVover 1	0.69	-0.1726	7.0e-02	0.014	*	-15.86
comorbidityCVover 1:followupTime		0.0122	1.9e-02	0.525		1.22
comorbidityDiabetestype1	0.58	-0.0527	2.1e-01	0.800		-5.13
comorbidityDiabetestype1:followupTime		0.0317	5.2e-02	0.544		3.22
comorbidityDiabetestype2	0.65	0.0450	5.0e-02	0.365		4.60
comorbidityDiabetestype2:followupTime		0.0080	1.3e-02	0.538		0.81
comorbidityGastrointestinal	0.60	0.1229	8.6e-02	0.154		13.08
comorbidityGastrointestinal:followupTime		-0.0217	2.0e-02	0.288		-2.15
CRP	0.63	0.0307	2.0e-02	0.129		3.12
CRP:followupTime		-0.0074	6.0e-03	0.221		-0.74
DBP	1.00	0.0059	1.4e-02	0.672		0.59
DBP:followupTime		0.0107	4.2e-03	0.012	*	1.07
familyHistoryIHD0	0.61	0.0449	6.5e-02	0.493		4.59
followupTime	0.63	-0.0448	2.3e-02	0.056		-4.38
Hb	0.96	0.0465	1.6e-02	0.004	**	4.76
Hb:followupTime		0.0053	6.0e-03	0.373		0.54
med.ACE.ARB		-0.0119	4.0e-02	0.764		-1.19
med.ACE.ARB:followupTime		0.0164	1.2e-02	0.162		1.66
med.AlphaBlockers	0.51	0.0054	3.9e-02	0.889		0.54
med.AlphaBlockers:followupTime		0.0341	1.3e-02	0.008	**	3.47
med.Diuretics	0.69	-0.0889	3.6e-02	0.015	*	-8.50
med.Diuretics:followupTime		0.0065	1.1e-02	0.538		0.66
med.Epo	1.00	-0.1739	5.0e-02	0.001	**	-15.96
med.Epo:followupTime		-0.0124	1.7e-02	0.470		-1.23
med.Iron	0.83	-0.0300	3.9e-02	0.444		-2.95
med.Iron:followupTime		-0.0183	1.1e-02	0.106		-1.81
med.ParenteralIron		0.0935	4.7e-02	0.049	*	9.80
med.ParenteralIron:followupTime		-0.0328	1.7e-02	0.059		-3.22
med.VitaminD	0.85	-0.0931	4.2e-02	0.027	*	-8.89
med.VitaminD:followupTime		-0.0153	1.6e-02	0.337		-1.52
numberAntihypertensives		0.0048	2.2e-02	0.828		0.48
numberAntihypertensives:followupTime		-0.0083	6.9e-03	0.229		-0.82
occupation0ManagerialProfessional	0.55	0.0590	8.3e-02	0.479		6.08
occupation0Intermediate	0.55	0.0224	9.1e-02	0.805		2.27
occupation0NeverWorkedUnemployed		0.3974	4.0e-01	0.319		48.80

Table 18: Standardised model summary for disease renovascular *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
PO	1.00	-0.0929	1.6e-02	0.000	***	-8.88
PO:followupTime		0.0049	5.0e-03	0.329		0.49
PP	0.87	0.0411	1.5e-02	0.005	**	4.19
PP:followupTime		-0.0032	4.7e-03	0.490		-0.32
PTH	0.96	-0.0522	2.0e-02	0.009	**	-5.09
PTH:followupTime		0.0075	7.3e-03	0.304		0.75
Pu	0.88	-0.0121	2.4e-02	0.620		-1.21
Pu:followupTime		-0.0056	4.6e-03	0.220		-0.56
smokingStatus0active	0.65	-0.2123	1.2e-01	0.074		-19.12
smokingStatus0ex-smoker	0.65	-0.1537	8.9e-02	0.087		-14.25
totalCholesterol	0.79	0.0076	1.7e-02	0.649		0.77
totalCholesterol:followupTime		-0.0153	5.6e-03	0.007	**	-1.52
weeklyAlcohol01 to 14	0.68	-0.0555	8.0e-02	0.491		-5.40
weeklyAlcohol0over 14	0.68	0.1141	9.5e-02	0.229		12.09

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.8 Unknown disease

Average effects

The key *average effects* are:

- Higher levels of Hb and total CO2 are associated with higher levels of eGFR. Note that poor kidney function may result in low levels of Hb and total CO2.
- Vitamin D supplements are associated with lower levels of eGFR. Poor kidney function is associated with vitamin D deficiency.
- Higher levels of PO and PTH are associated with lower levels of eGFR. Poor kidney function is often associated with higher levels of these biochemicals.
- Older age at baseline is associated with a lower level of eGFR.

Weaker associations with a significance level of 0.01-0.05 are:

- Higher levels of Pu are associated with lower levels of eGFR. Poor kidney function is often associated with higher levels of protein in the urine.
- DBP is associated with a less rapid decline in kidney function. This is unexpected because as PP rises, for example with older age and worsening kidney function, the clinicians expect DBP to fall.
- Higher numbers of AKI episodes are associated with higher levels of eGFR. This is an anomalous result given AKI would typically be associated with poor kidney function. However this disease group *unknown* is heterogeneous so perhaps the results are being skewed by a sub-group who are more susceptible to AKI.

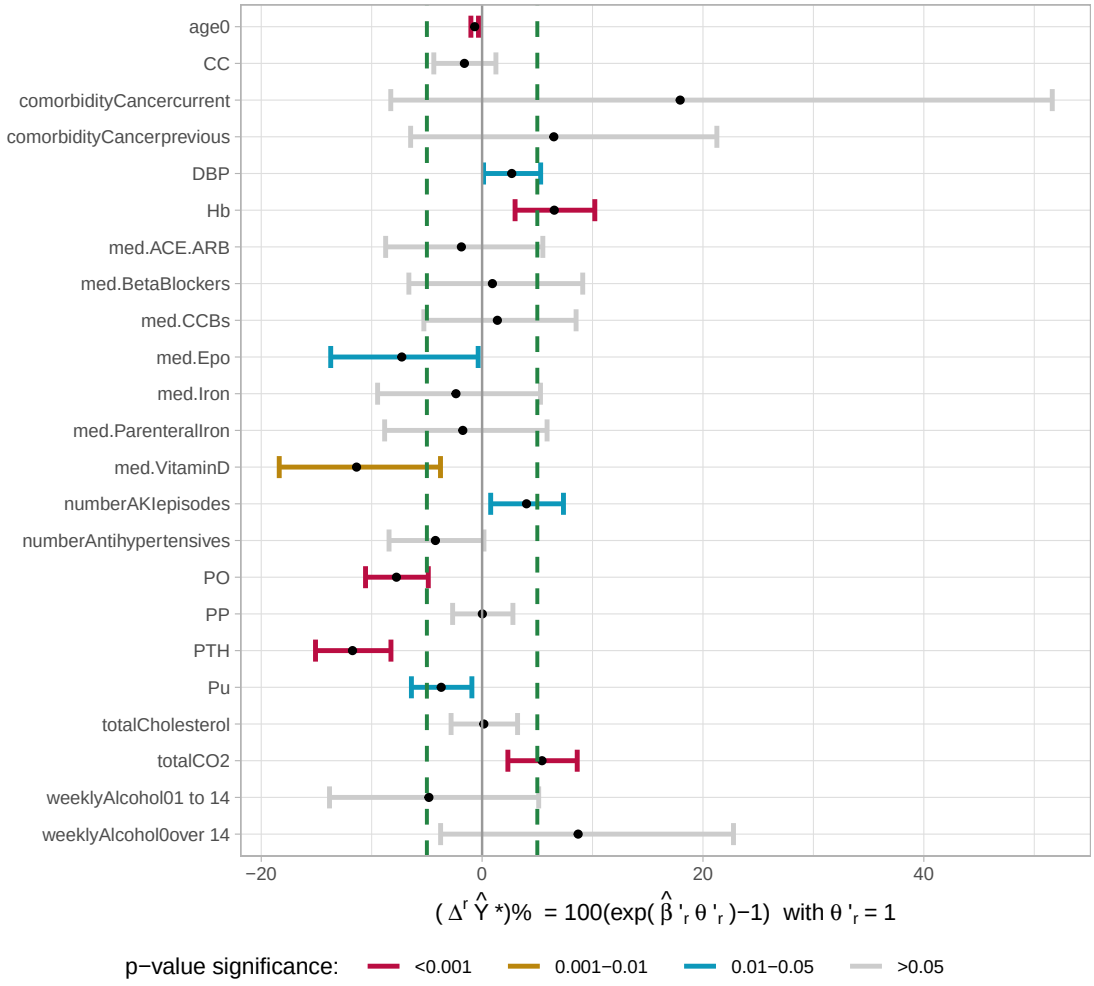

Figure 34: Average effects - relative change in eGFR for standardised model using 95% CIs: unknown

Temporal effects

The key *temporal effects* are:

- PO is associated with a more rapid decline in eGFR.
- PTH is associated with slower progression of kidney disease. This result is counterintuitive as higher levels of PTH tend to occur with worsening kidney function.

There is also a relatively weak association between the number of antihypertensives and slower progression. However given the heterogeneous nature of this disease group it is perhaps plausible that a sub-group has their progression slowed by these drugs.

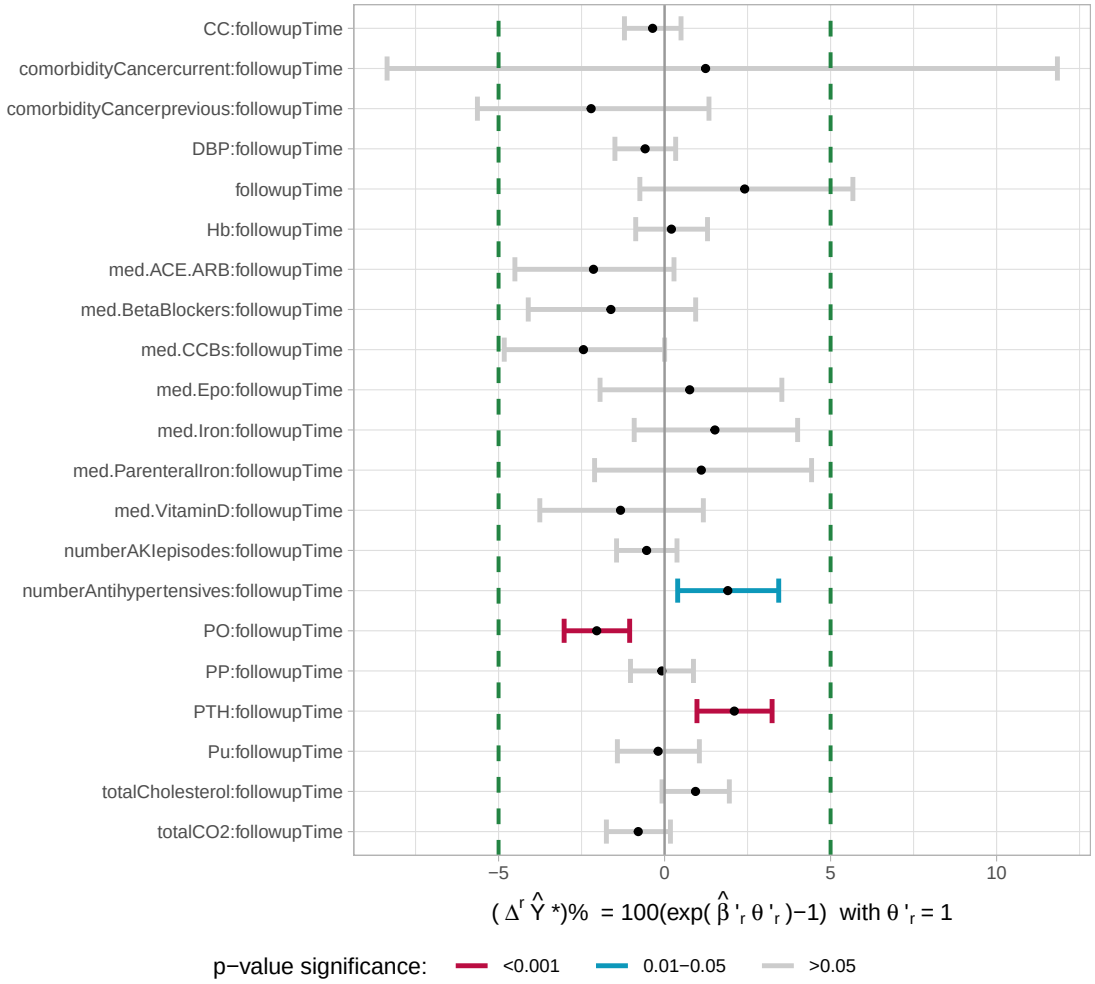

Figure 35: Temporal effects - relative change in eGFR for standardised model using 95% CIs: unknown

Parameter values

Table 19: Standardised model summary for disease unknown

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	3.8130	1.4e-01	0.000	***	
age0	0.86	-0.0067	1.8e-03	0.000	***	-0.67
CC	0.57	-0.0162	1.5e-02	0.282		-1.60
CC:followupTime		-0.0036	4.5e-03	0.422		-0.36
comorbidityCancercurrent	0.55	0.1650	1.3e-01	0.211		17.94
comorbidityCancercurrent:followupTime		0.0123	5.2e-02	0.814		1.23
comorbidityCancerprevious	0.55	0.0629	6.8e-02	0.356		6.49
comorbidityCancerprevious:followupTime		-0.0224	1.9e-02	0.232		-2.21
DBP		0.0265	1.3e-02	0.045	*	2.68
DBP:followupTime		-0.0059	4.8e-03	0.225		-0.59
followupTime		0.0238	1.6e-02	0.147		2.41
Hb	0.99	0.0633	1.8e-02	0.000	***	6.53
Hb:followupTime		0.0021	5.7e-03	0.716		0.21
med.ACE.ARB	0.55	-0.0189	3.8e-02	0.619		-1.88
med.ACE.ARB:followupTime		-0.0216	1.3e-02	0.092		-2.14
med.BetaBlockers	0.54	0.0093	4.1e-02	0.820		0.93
med.BetaBlockers:followupTime		-0.0163	1.3e-02	0.226		-1.62
med.CCBs	0.70	0.0137	3.6e-02	0.700		1.38
med.CCBs:followupTime		-0.0247	1.3e-02	0.057		-2.44
med.Epo	0.85	-0.0755	3.8e-02	0.046	*	-7.27
med.Epo:followupTime		0.0075	1.4e-02	0.597		0.76
med.Iron		-0.0240	4.0e-02	0.545		-2.37
med.Iron:followupTime		0.0151	1.3e-02	0.237		1.52
med.ParenteralIron	0.67	-0.0176	3.9e-02	0.653		-1.75
med.ParenteralIron:followupTime		0.0110	1.7e-02	0.516		1.11
med.VitaminD	0.99	-0.1207	4.3e-02	0.005	**	-11.37
med.VitaminD:followupTime		-0.0134	1.3e-02	0.308		-1.33
numberAKIepisodes	0.63	0.0394	1.7e-02	0.018	*	4.02
numberAKIepisodes:followupTime		-0.0054	4.8e-03	0.259		-0.54
numberAntihypertensives		-0.0432	2.4e-02	0.067		-4.23
numberAntihypertensives:followupTime		0.0189	7.8e-03	0.016	*	1.91
PO	1.00	-0.0808	1.6e-02	0.000	***	-7.76
PO:followupTime		-0.0206	5.3e-03	0.000	***	-2.04
PP		0.0002	1.4e-02	0.987		0.02
PP:followupTime		-0.0008	5.0e-03	0.868		-0.08
PTH	0.98	-0.1248	2.0e-02	0.000	***	-11.74
PTH:followupTime		0.0208	5.8e-03	0.000	***	2.10
Pu	0.82	-0.0377	1.5e-02	0.012	*	-3.70
Pu:followupTime		-0.0019	6.5e-03	0.765		-0.19
totalCholesterol	0.58	0.0016	1.6e-02	0.922		0.16
totalCholesterol:followupTime		0.0093	5.3e-03	0.079		0.93

Table 19: Standardised model summary for disease unknown *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
totalCO2	1.00	0.0528	1.6e-02	0.001	**	5.43
totalCO2:followupTime		-0.0080	5.1e-03	0.119		-0.79
weeklyAlcohol01 to 14	0.51	-0.0494	5.2e-02	0.343		-4.82
weeklyAlcohol0over 14	0.51	0.0834	6.4e-02	0.192		8.70

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.3.9 Single model all diseases

As shown in Table 20 the model for this category has disease as an explanatory variable. As this explanatory variable is not directly of interest we do not include it in Figures 36 and 37.

Average effects

The key *average effects* are:

- High levels of the biochemicals CC, PO and PTH are associated with lower levels of eGFR.
- A higher count of cardiovascular (CV) diseases is associated with lower levels of eGFR.
- Higher levels of Hb and total CO2 are associated with higher levels of eGFR. (Note that poor kidney function may result in low levels of Hb and total CO2.)
- The ACE inhibitors and ARBs (med.ACE.ARB) are associated with higher levels of eGFR.
- The treatments EPO, diuretics and vitamin D are associated with lower levels of eGFR
- Older age at baseline is associated with a lower level of eGFR.

Weaker associations at the 0.01-0.05 significance level are:

- The treatments CCBs and 'other medications' are associated with lower levels of eGFR.
- Higher values of PP are associated with higher levels of eGFR. This result is counterintuitive as it is expected that PP will increase with age, worsening kidney function and poorer cardiovascular health.
- Higher levels of total cholesterol are associated with higher levels of eGFR. It is unclear why this should be the case when higher cholesterol is typically associated with poorer health.
- Higher values of DBP are associated with higher levels of eGFR. This is not expected as DBP decreases with both age and increased PP.

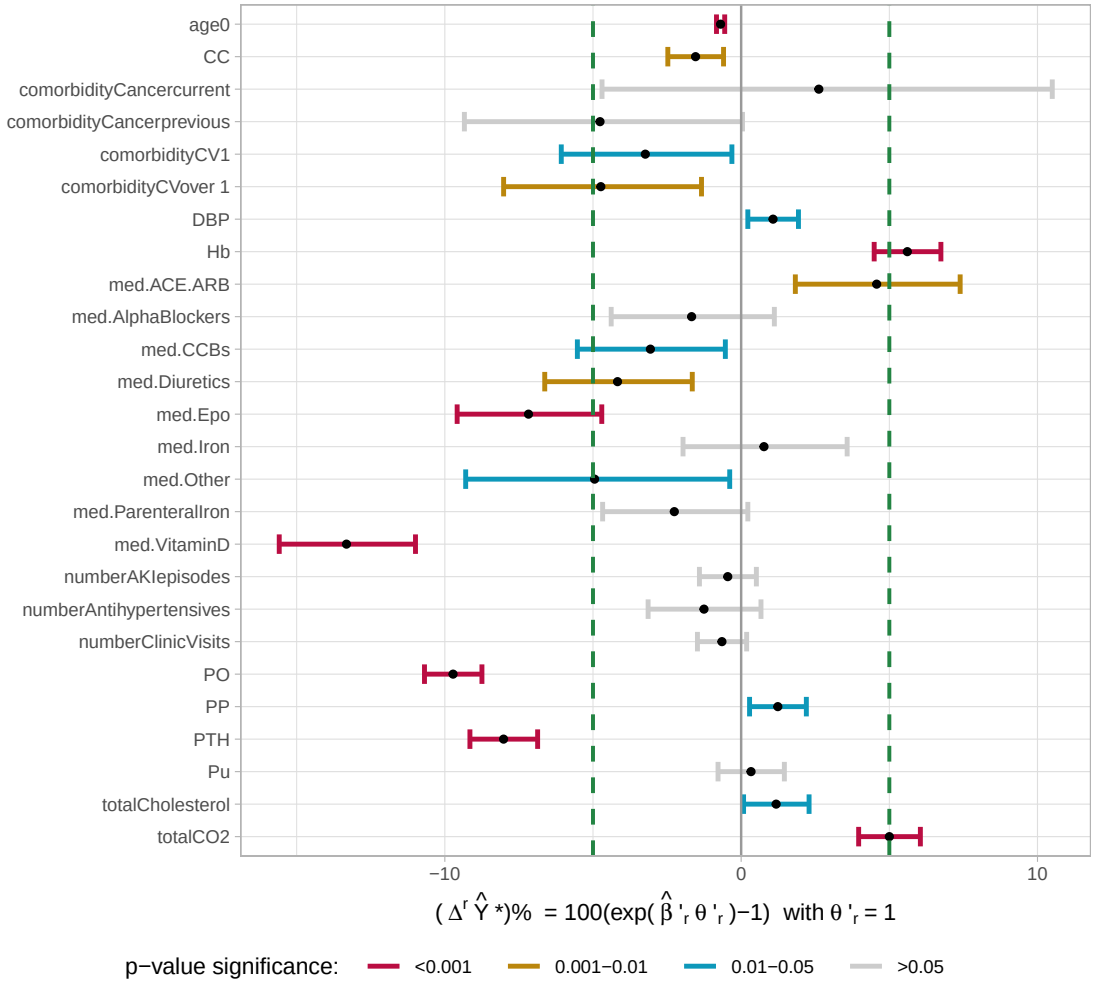

Figure 36: Average effects - relative change in eGFR for standardised model using 95% CIs: single model all diseases

Temporal effects

The key *temporal effects* are:

- A negative value of follow-up time indicates that eGFR is falling off over time. The dominance of follow-up time may indicate that our model and the SKS dataset are missing at least one key factor.
- Hb is associated with a less rapid decline in eGFR.
- PO, Pu and total CO2 are associated in a more rapid decline in eGFR.
- PTH is associated with a less rapid decline in eGFR. This is an unexpected result as the risk of increased PTH is associated with poorer kidney function.

We also note that iron taken orally is relatively weakly associated with a more rapid decline in eGFR. As kidney function reduces there is an increased likelihood of anaemia. One treatment option for this condition is iron taken orally.

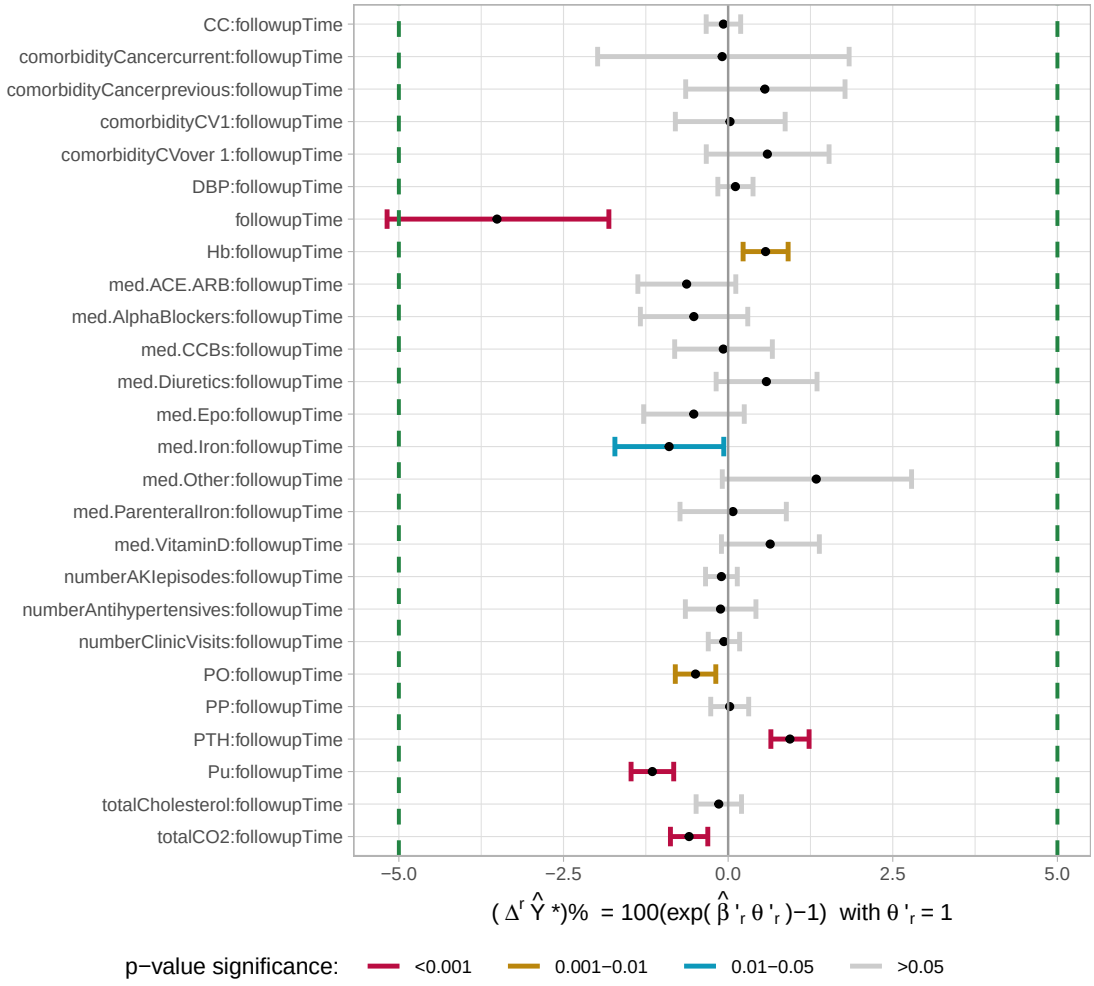

Figure 37: Temporal effects - relative change in eGFR for standardised model using 95% CIs: single model all diseases

Parameter values

Table 20: Standardised model summary for single model all diseases

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
(Intercept)	1.00	3.9816	5.5e-02	0.000	***	
age0	1.00	-0.0070	7.1e-04	0.000	***	-0.69
CC	0.89	-0.0155	4.9e-03	0.002	**	-1.54
CC:followupTime		-0.0007	1.3e-03	0.592		-0.07
comorbidityCancercurrent	0.72	0.0259	3.8e-02	0.495		2.62
comorbidityCancercurrent:followupTime		-0.0009	9.8e-03	0.926		-0.09
comorbidityCancerprevious	0.72	-0.0488	2.5e-02	0.053		-4.76
comorbidityCancerprevious:followupTime		0.0056	6.2e-03	0.367		0.56
comorbidityCV1	0.80	-0.0329	1.5e-02	0.031	*	-3.23
comorbidityCV1:followupTime		0.0003	4.3e-03	0.947		0.03
comorbidityCVover 1	0.80	-0.0485	1.8e-02	0.007	**	-4.74
comorbidityCVover 1:followupTime		0.0059	4.7e-03	0.211		0.60
DBP	1.00	0.0107	4.3e-03	0.013	*	1.08
DBP:followupTime		0.0011	1.4e-03	0.419		0.11
disease diabetic nephropathy	1.00	-0.0963	3.2e-02	0.003	**	-9.18
disease glomerulonephritis	1.00	0.0091	3.4e-02	0.788		0.91
disease HKD	1.00	-0.0536	3.5e-02	0.124		-5.22
disease obstruction	1.00	-0.3433	8.5e-02	0.000	***	-29.05
disease polycystic kidney disease	1.00	-0.1782	4.4e-02	0.000	***	-16.32
disease pyelonephritis	1.00	-0.1395	4.4e-02	0.001	**	-13.02
disease renovascular disease	1.00	-0.0197	4.1e-02	0.633		-1.95
disease unknown	1.00	-0.0556	3.5e-02	0.110		-5.41
followupTime	1.00	-0.0357	8.9e-03	0.000	***	-3.51
Hb	1.00	0.0546	5.5e-03	0.000	***	5.61
Hb:followupTime		0.0057	1.7e-03	0.001	**	0.57
med.ACE.ARB	0.97	0.0447	1.4e-02	0.001	**	4.57
med.ACE.ARB:followupTime		-0.0063	3.8e-03	0.099		-0.63
med.AlphaBlockers	0.83	-0.0169	1.4e-02	0.240		-1.67
med.AlphaBlockers:followupTime		-0.0052	4.2e-03	0.214		-0.52
med.CCBs	0.72	-0.0311	1.3e-02	0.018	*	-3.06
med.CCBs:followupTime		-0.0007	3.8e-03	0.847		-0.07
med.Diuretics	0.85	-0.0426	1.3e-02	0.001	**	-4.17
med.Diuretics:followupTime		0.0058	3.9e-03	0.138		0.58
med.Epo	1.00	-0.0745	1.3e-02	0.000	***	-7.18
med.Epo:followupTime		-0.0052	3.9e-03	0.184		-0.52
med.Iron	0.54	0.0077	1.4e-02	0.586		0.77
med.Iron:followupTime		-0.0090	4.3e-03	0.035	*	-0.90
med.Other	0.58	-0.0507	2.4e-02	0.035	*	-4.95
med.Other:followupTime		0.0133	7.3e-03	0.067		1.34
med.ParenteralIron	0.71	-0.0228	1.3e-02	0.076		-2.26
med.ParenteralIron:followupTime		0.0007	4.1e-03	0.860		0.07

Table 20: Standardised model summary for single model all diseases *(continued)*

parameter	prop[a]	$\hat{\beta}'_r$	se	p-value	stars[b]	$(\Delta^r \hat{Y}^*)\%$ [c]
med.VitaminD	1.00	-0.1429	1.4e-02	0.000	***	-13.32
med.VitaminD:followupTime		0.0064	3.8e-03	0.093		0.64
numberAKIepisodes	0.77	-0.0045	4.9e-03	0.360		-0.45
numberAKIepisodes:followupTime		-0.0010	1.2e-03	0.408		-0.10
numberAntihypertensives	0.62	-0.0126	9.9e-03	0.202		-1.25
numberAntihypertensives:followupTime		-0.0012	2.7e-03	0.675		-0.12
numberClinicVisits	0.99	-0.0065	4.3e-03	0.129		-0.65
numberClinicVisits:followupTime		-0.0006	1.2e-03	0.599		-0.06
PO	1.00	-0.1023	5.5e-03	0.000	***	-9.72
PO:followupTime		-0.0050	1.6e-03	0.002	**	-0.50
PP	0.92	0.0123	4.8e-03	0.011	*	1.24
PP:followupTime		0.0002	1.5e-03	0.874		0.02
PTH	1.00	-0.0836	6.4e-03	0.000	***	-8.02
PTH:followupTime		0.0093	1.5e-03	0.000	***	0.94
Pu	0.93	0.0033	5.7e-03	0.561		0.33
Pu:followupTime		-0.0116	1.7e-03	0.000	***	-1.15
totalCholesterol	0.62	0.0117	5.6e-03	0.036	*	1.18
totalCholesterol:followupTime		-0.0014	1.8e-03	0.421		-0.14
totalCO2	1.00	0.0488	5.1e-03	0.000	***	5.00
totalCO2:followupTime		-0.0059	1.5e-03	0.000	***	-0.59

[a] proportion of bootstraps in which variable was selected

[b] p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *

[c] $(\Delta^r \hat{Y}^*)\% = 100(\exp(\hat{\beta}'_r \theta'_r) - 1)$ with $\theta'_r = 1$

7.4 Rates of change over time

Here we estimate the average, population level, rates of change over time using the time derivative method described in Section 4.2. We use the explanatory variables selected in section 5 but do not standardise them when fitting the LME model described by Equation 2 in Section 3. The estimated regression coefficients are used when computing the expected rates over time. Note that we do not fit the time derivative of the LME model. All rates of change estimates are based on patients with more than 2 follow-up records. For a given variable the average rates over time for each individual are computed, after which we use bootstrapping to obtain summary statistics; we resample with replacement 2,000 times. The 95% confidence intervals are calculated using a non-parametric bootstrap confidence interval method. Specifically, they are computed using the adjusted bootstrap percentile method (in particular the bias-corrected and accelerated method, BCa) provided by the boot::boot.ci() R-function, for details see (83,84). In Sections 7.4.2 to 7.4.11 a quantity is significant if these confidence intervals do not cover zero.

First, in Section 7.4.1 we give the expected rate of change of the outcome variable eGFR for each disease category. This is estimated with $\mathbb{E}(\dot{\hat{Y}}_i^*)$ as described in Section 4.2.2. These results constitute our main findings regarding rates. Secondly, in Sections 7.4.2 to 7.4.11 we show the breakdown of the estimated rates for each disease category. In particular we compute $\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ with Equation 23. These results are shown in Figures 39 to 47 and tabulated in Tables 22 to 30. These results are supplementary to those reported in Section 7.4.1.

7.4.1 Overall average rate of decline for each disease

Figure 38 shows the average annual rate of decline in eGFR for each disease category; details are given in Table 21. We consider all these rates to be significant in so much as the confidence intervals do not cover zero. As shown the entire cohort, 'single model all diseases' labelled 'All', is on average loosing eGFR at a rate of -1.1 mL/min/1.73m^2/year. In contrast PKD has the highest rate at -3.5 mL/min/1.73m^2/year.

In 2013, using the SKS data, Hoefield (85) reported that PKD and diabetic nephropathy patients exhibited on average a 2.7 ($\pm$ 0.3) and 0.7 ($\pm$ 0.3) ml/min/1.73m^2/year faster rate of decline in eGFR, respectively, compared to patients with glomerulonephritis. In these terms, with our model, we report patients with PKD and diabetic nephropathy have on average a 2.5 ($\pm$ 0.1) and 0.6 ($\pm$ 0.05) ml/min/1.73m^2/year faster rate of decline in eGFR, respectively, compared to those with glomerulonephritis. Given the whole cohort Hoefield (85) reports a median of -1.2 (IQR: -3.6, 0.2) ml/min/1.73m^2/year whereas we report a median of -0.9 (IQR: -2.1, 0.1) mL/min/1.73m^2/year. Furthermore our estimates for average annual rates of decline of diabetic nephropathy and PKD, 1.5 mL/min/1.73m^2/year and 3.5 mL/min/1.73m^2/year, respectively, are similar to those reported elsewhere. For example (32) reports the average rate of decline for

diabetic nephropathy as 1.7 mL/min/1.73m^2/year (32), and (33) reports ~3 mL/min/1.73m^2/year for PKD. Clearly our model gives similar estimates to those reported previously.

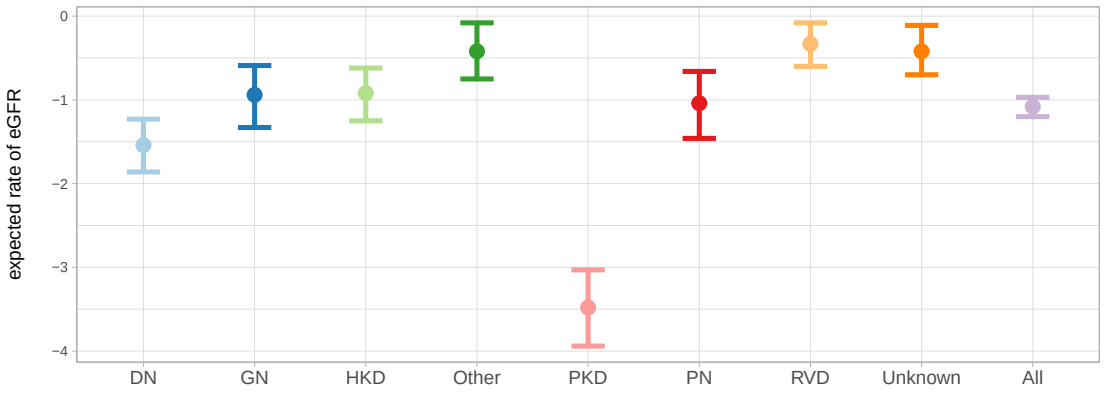

Figure 38: Estimated rate of decline in eGFR by disease

Table 21: Summary for rates of change in eGFR across all diseases

disease category	$\mathbb{E}(\dot{\hat{Y}}_i^*)$	CI
diabetic nephropathy	-1.54	(-1.86,-1.23)
glomerulonephritis	-0.94	(-1.33,-0.59)
HKD	-0.92	(-1.25,-0.62)
other	-0.42	(-0.75,-0.08)
PKD	-3.48	(-3.94,-3.03)
pyelonephritis	-1.04	(-1.46,-0.66)
renovascular disease	-0.33	(-0.60,-0.08)
unknown	-0.42	(-0.70,-0.11)
single model all diseases	-1.08	(-1.20,-0.97)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^*)$ has units mL/min/1.73m^2/year

7.4.2 Diabetic nephropathy

On average across the population the dominant terms:

- bodyMassIndex:followupTime, CC:followupTime, Hb:followupTime, PO:followupTime contribute to a less rapid decline in kidney function; i.e. terms have positive slope.
- DBP:followupTime, med.ACE.ARB:followupTime, numberAntihypertensives:followupTime, PP:followupTime, PTH and Pu:followupTime contribute to a more rapid decline in kidney function; i.e. terms have negative slope.

Note that each term is comprised of a regression parameter multiplied by the averaged time derivative of the corresponding explanatory variable, see Equation 23.

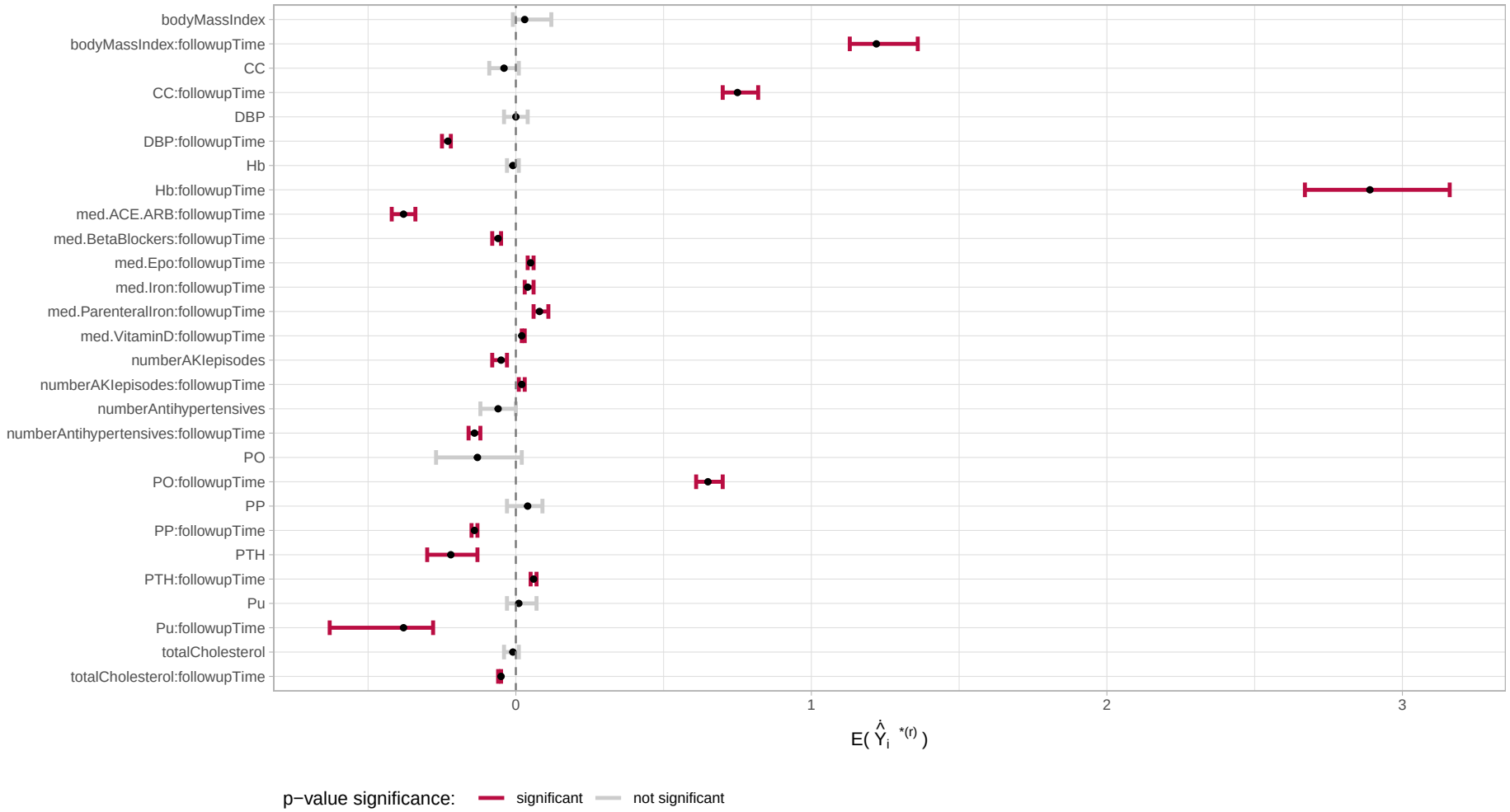

Figure 39: Rate estimates with 95% CIs for diabetic nephropathy

Table 22: Estimated average rate of change over time for disease diabetic nephropathy

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-1.54	(-1.86,-1.23)
biochemical		
CC	-0.04	(-0.09,0.01)
CC:followupTime	0.75	(0.70,0.82)
Hb	-0.01	(-0.03,0.01)
Hb:followupTime	2.89	(2.67,3.16)
PO	-0.13	(-0.27,0.02)
PO:followupTime	0.65	(0.61,0.70)
PTH	-0.22	(-0.30,-0.13)
PTH:followupTime	0.06	(0.05,0.07)
Pu	0.01	(-0.03,0.07)
Pu:followupTime	-0.38	(-0.63,-0.28)
totalCholesterol	-0.01	(-0.04,0.01)
totalCholesterol:followupTime	-0.05	(-0.06,-0.05)
catagorical		
med.ACE.ARB:followupTime	-0.38	(-0.42,-0.34)
med.BetaBlockers:followupTime	-0.06	(-0.08,-0.05)
med.Epo:followupTime	0.05	(0.04,0.06)
med.Iron:followupTime	0.04	(0.03,0.06)
med.ParenteralIron:followupTime	0.08	(0.06,0.11)
med.VitaminD:followupTime	0.02	(0.02,0.03)
general		
bodyMassIndex	0.03	(-0.01,0.12)
bodyMassIndex:followupTime	1.22	(1.13,1.36)
DBP	0.00	(-0.04,0.04)
DBP:followupTime	-0.23	(-0.25,-0.22)
numberAKIepisodes	-0.05	(-0.08,-0.03)
numberAKIepisodes:followupTime	0.02	(0.01,0.03)
numberAntihypertensives	-0.06	(-0.12,0.00)
numberAntihypertensives:followupTime	-0.14	(-0.16,-0.12)
PP	0.04	(-0.03,0.09)
PP:followupTime	-0.14	(-0.15,-0.13)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.3 Glomerulonephritis

On average across the population the dominant terms:

- DBP:followupTime, Hb:followupTime and PTH:followupTime contribute to a less rapid decline in kidney function.
- med.ACE.ARB:followupTime, PO:followupTime, PTH, Pu:followupTime and totalCO2:followupTime contribute to a more rapid decline in kidney function.

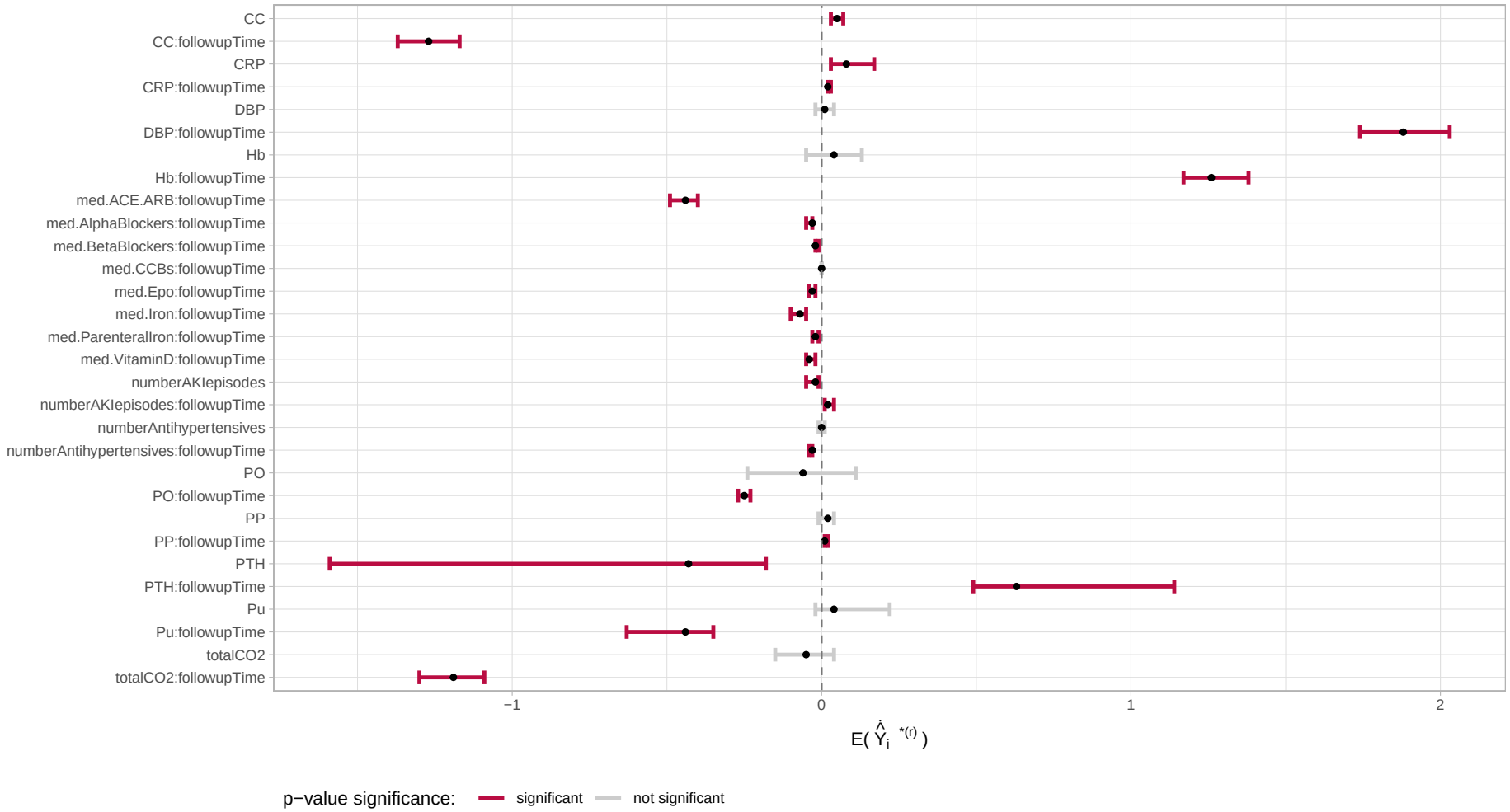

Figure 40: Rate estimates with 95% CIs for glomerulonephritis

Table 23: Estimated average rate of change over time for disease glomerulonephritis

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-0.94	(-1.33,-0.59)
biochemical		
CC	0.05	(0.03,0.07)
CC:followupTime	-1.27	(-1.37,-1.17)
CRP	0.08	(0.03,0.17)
CRP:followupTime	0.02	(0.02,0.03)
Hb	0.04	(-0.05,0.13)
Hb:followupTime	1.26	(1.17,1.38)
PO	-0.06	(-0.24,0.11)
PO:followupTime	-0.25	(-0.27,-0.23)
PTH	-0.43	(-1.59,-0.18)
PTH:followupTime	0.63	(0.49,1.14)
Pu	0.04	(-0.02,0.22)
Pu:followupTime	-0.44	(-0.63,-0.35)
totalCO2	-0.05	(-0.15,0.04)
totalCO2:followupTime	-1.19	(-1.30,-1.09)
catagorical		
med.ACE.ARB:followupTime	-0.44	(-0.49,-0.40)
med.AlphaBlockers:followupTime	-0.03	(-0.05,-0.03)
med.BetaBlockers:followupTime	-0.02	(-0.02,-0.01)
med.CCBs:followupTime	0.00	(0.00,0.00)
med.Epo:followupTime	-0.03	(-0.04,-0.02)
med.Iron:followupTime	-0.07	(-0.10,-0.05)
med.ParenteralIron:followupTime	-0.02	(-0.03,-0.01)
med.VitaminD:followupTime	-0.04	(-0.05,-0.02)
general		
DBP	0.01	(-0.02,0.04)
DBP:followupTime	1.88	(1.74,2.03)
numberAKIepisodes	-0.02	(-0.05,-0.01)
numberAKIepisodes:followupTime	0.02	(0.01,0.04)
numberAntihypertensives	0.00	(-0.01,0.01)
numberAntihypertensives:followupTime	-0.03	(-0.04,-0.03)
PP	0.02	(-0.01,0.04)
PP:followupTime	0.01	(0.01,0.02)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.4 Hypertensive kidney disease

On average across the population the dominant terms:

- CC:followupTime, Hb:followupTime, PP:followupTime and numberAntihypertensives:followupTime contribute to a less rapid decline in kidney function.
- numberClinicVisits:followupTime, PO:followupTime, Pu:followupTime and totalCO2:followupTime contribute to a more rapid decline in kidney function.

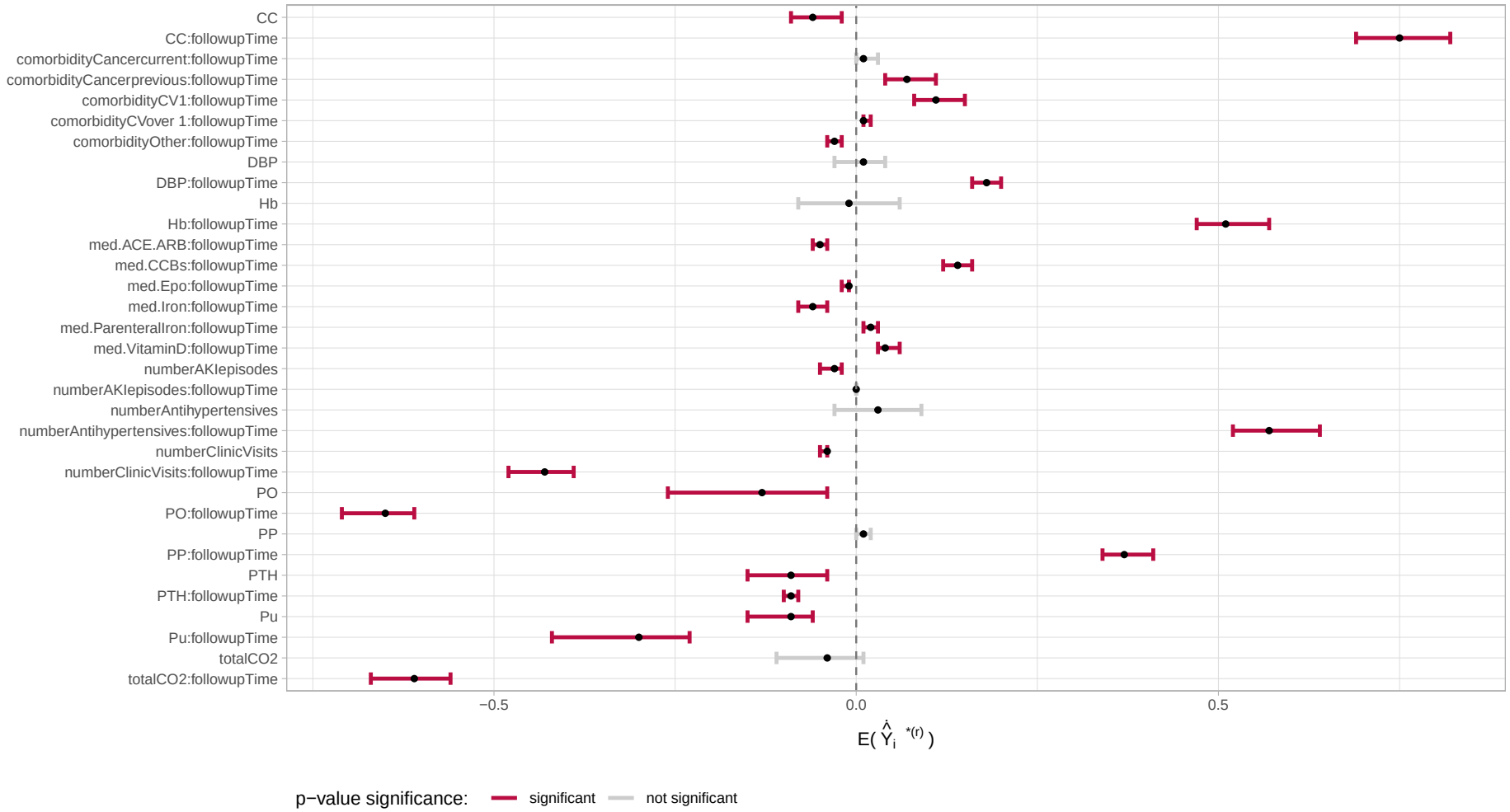

Figure 41: Rate estimates with 95% CIs for HKD

Table 24: Estimated average rate of change over time for disease HKD

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-0.92	(-1.25,-0.62)
biochemical		
CC	-0.06	(-0.09,-0.02)
CC:followupTime	0.75	(0.69,0.82)
Hb	-0.01	(-0.08,0.06)
Hb:followupTime	0.51	(0.47,0.57)
PO	-0.13	(-0.26,-0.04)
PO:followupTime	-0.65	(-0.71,-0.61)
PTH	-0.09	(-0.15,-0.04)
PTH:followupTime	-0.09	(-0.10,-0.08)
Pu	-0.09	(-0.15,-0.06)
Pu:followupTime	-0.30	(-0.42,-0.23)
totalCO2	-0.04	(-0.11,0.01)
totalCO2:followupTime	-0.61	(-0.67,-0.56)
catagorical		
comorbidityCancercurrent:followupTime	0.01	(0.00,0.03)
comorbidityCancerprevious:followupTime	0.07	(0.04,0.11)
comorbidityCV1:followupTime	0.11	(0.08,0.15)
comorbidityCVover 1:followupTime	0.01	(0.01,0.02)
comorbidityOther:followupTime	-0.03	(-0.04,-0.02)
med.ACE.ARB:followupTime	-0.05	(-0.06,-0.04)
med.CCBs:followupTime	0.14	(0.12,0.16)
med.Epo:followupTime	-0.01	(-0.02,-0.01)
med.Iron:followupTime	-0.06	(-0.08,-0.04)
med.ParenteralIron:followupTime	0.02	(0.01,0.03)
med.VitaminD:followupTime	0.04	(0.03,0.06)
general		
DBP	0.01	(-0.03,0.04)
DBP:followupTime	0.18	(0.16,0.20)
numberAKIepisodes	-0.03	(-0.05,-0.02)
numberAKIepisodes:followupTime	0.00	(0.00,0.00)
numberAntihypertensives	0.03	(-0.03,0.09)
numberAntihypertensives:followupTime	0.57	(0.52,0.64)
numberClinicVisits	-0.04	(-0.05,-0.04)
numberClinicVisits:followupTime	-0.43	(-0.48,-0.39)
PP	0.01	(0.00,0.02)
PP:followupTime	0.37	(0.34,0.41)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.5 Other

On average across the population the dominant terms:

- CC:followupTime, numberClinicVisits, PTH:followupTime and totalCO2:followupTime contribute to a less rapid decline in kidney function.
- Hb:followupTime and PO:followupTime contribute to a more rapid decline in kidney function.

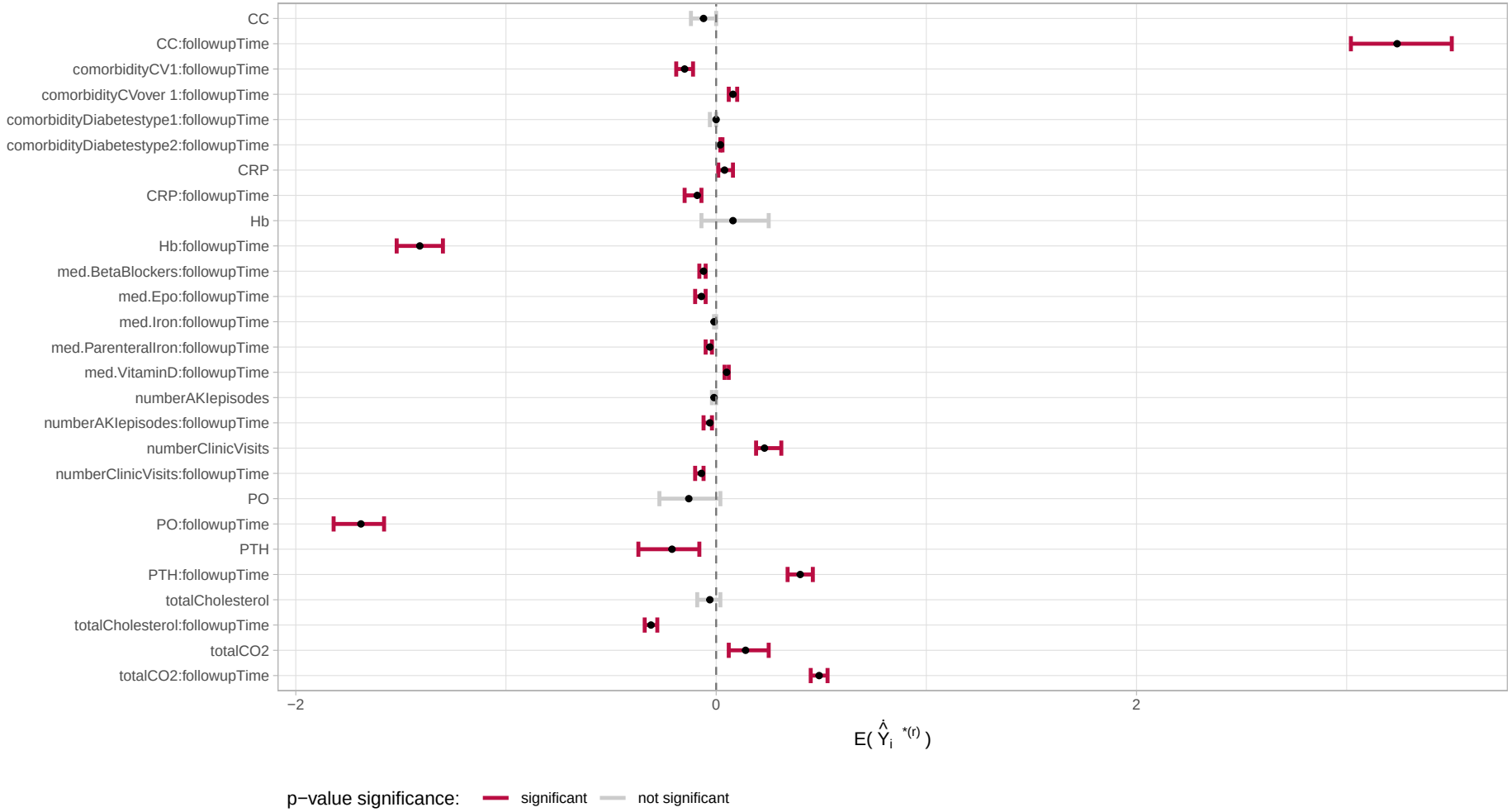

Figure 42: Rate estimates with 95% CIs for disease other

Table 25: Estimated average rate of change over time for disease other

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-0.42	(-0.75,-0.08)
biochemical		
CC	-0.06	(-0.12,0.00)
CC:followupTime	3.24	(3.02,3.50)
CRP	0.04	(0.01,0.08)
CRP:followupTime	-0.09	(-0.15,-0.07)
Hb	0.08	(-0.07,0.25)
Hb:followupTime	-1.41	(-1.52,-1.30)
PO	-0.13	(-0.27,0.02)
PO:followupTime	-1.69	(-1.82,-1.58)
PTH	-0.21	(-0.37,-0.08)
PTH:followupTime	0.40	(0.34,0.46)
totalCholesterol	-0.03	(-0.09,0.02)
totalCholesterol:followupTime	-0.31	(-0.34,-0.28)
totalCO2	0.14	(0.06,0.25)
totalCO2:followupTime	0.49	(0.45,0.53)
catagorical		
comorbidityCV1:followupTime	-0.15	(-0.19,-0.11)
comorbidityCVover 1:followupTime	0.08	(0.06,0.10)
comorbidityDiabetestype1:followupTime	0.00	(-0.03,0.00)
comorbidityDiabetestype2:followupTime	0.02	(0.02,0.03)
med.BetaBlockers:followupTime	-0.06	(-0.08,-0.05)
med.Epo:followupTime	-0.07	(-0.10,-0.05)
med.Iron:followupTime	-0.01	(-0.01,0.00)
med.ParenteralIron:followupTime	-0.03	(-0.05,-0.02)
med.VitaminD:followupTime	0.05	(0.04,0.06)
general		
numberAKIepisodes	-0.01	(-0.02,0.00)
numberAKIepisodes:followupTime	-0.03	(-0.06,-0.02)
numberClinicVisits	0.23	(0.19,0.31)
numberClinicVisits:followupTime	-0.07	(-0.10,-0.06)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.6 PKD

On average across the population the dominant terms:

- Hb:followupTime and numberAntihypertensives:followupTime contribute to a less rapid decline in kidney function.
- CC:followupTime, PO:followupTime and totalCO2:followupTime contribute to a more rapid decline in kidney function.

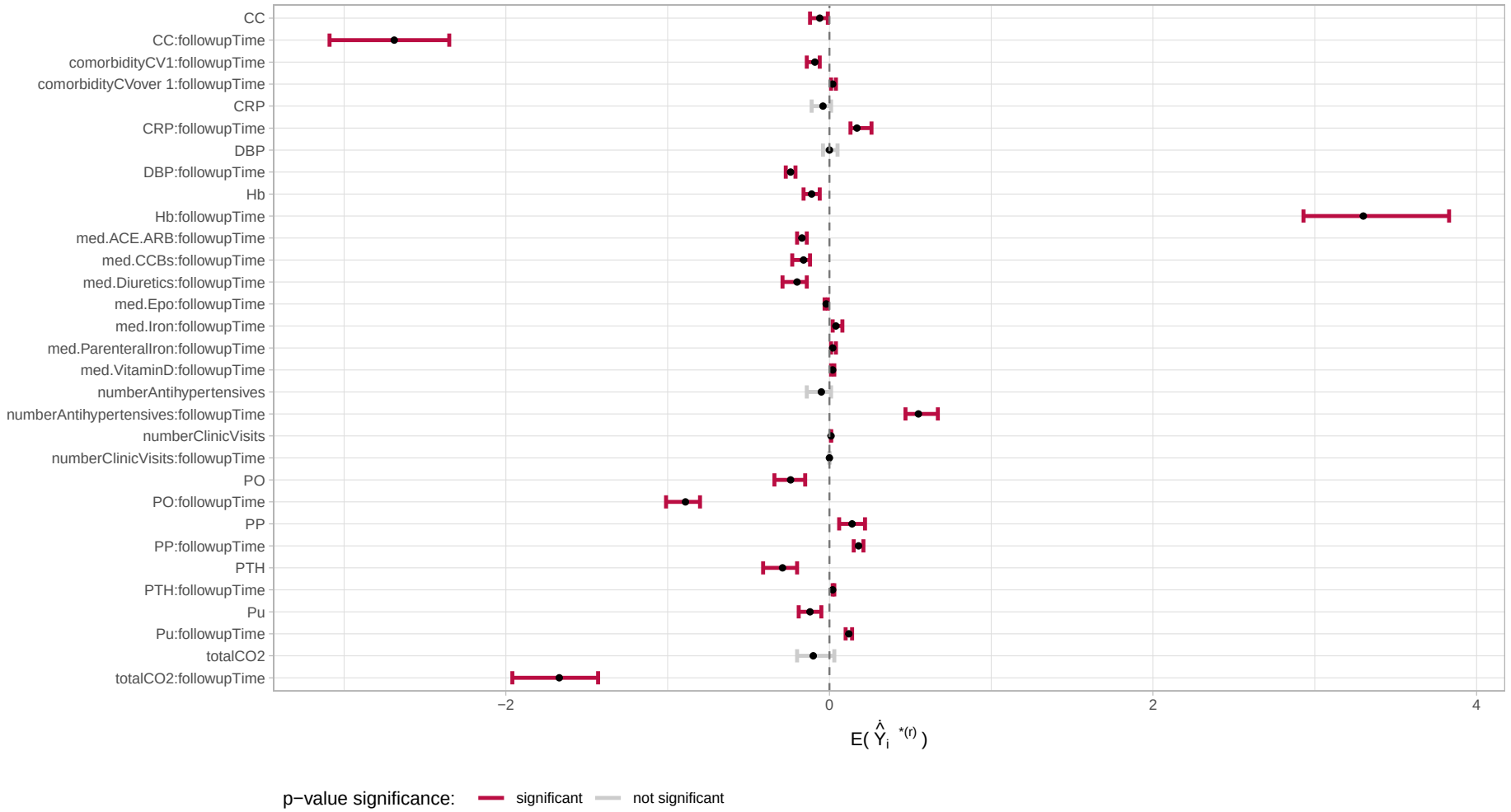

Figure 43: Rate estimates with 95% CIs for PKD

Table 26: Estimated average rate of change over time for disease PKD

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-3.48	(-3.94,-3.03)
biochemical		
CC	-0.06	(-0.12,-0.01)
CC:followupTime	-2.69	(-3.09,-2.35)
CRP	-0.04	(-0.11,0.01)
CRP:followupTime	0.17	(0.13,0.26)
Hb	-0.11	(-0.16,-0.06)
Hb:followupTime	3.30	(2.93,3.83)
PO	-0.24	(-0.34,-0.15)
PO:followupTime	-0.89	(-1.01,-0.80)
PTH	-0.29	(-0.41,-0.20)
PTH:followupTime	0.02	(0.02,0.03)
Pu	-0.12	(-0.19,-0.05)
Pu:followupTime	0.12	(0.10,0.14)
totalCO2	-0.10	(-0.20,0.03)
totalCO2:followupTime	-1.67	(-1.96,-1.43)
catagorical		
comorbidityCV1:followupTime	-0.09	(-0.14,-0.06)
comorbidityCVover 1:followupTime	0.02	(0.01,0.04)
med.ACE.ARB:followupTime	-0.17	(-0.20,-0.14)
med.CCBs:followupTime	-0.16	(-0.23,-0.12)
med.Diuretics:followupTime	-0.20	(-0.29,-0.14)
med.Epo:followupTime	-0.02	(-0.03,-0.01)
med.Iron:followupTime	0.04	(0.02,0.08)
med.ParenteralIron:followupTime	0.02	(0.01,0.04)
med.VitaminD:followupTime	0.02	(0.01,0.03)
general		
DBP	0.00	(-0.04,0.05)
DBP:followupTime	-0.24	(-0.27,-0.21)
numberAntihypertensives	-0.05	(-0.14,0.01)
numberAntihypertensives:followupTime	0.55	(0.47,0.67)
numberClinicVisits	0.01	(0.01,0.01)
numberClinicVisits:followupTime	0.00	(0.00,0.00)
PP	0.14	(0.06,0.22)
PP:followupTime	0.18	(0.15,0.21)

Note:

$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.7 Pyelonephritis

On average across the population the dominant terms:

- bodyMassIndex:followupTime, Hb:followupTime and PP:followupTime contribute to a less rapid decline in kidney function.
- totalCO2:followupTime contributes to a more rapid decline in kidney function.

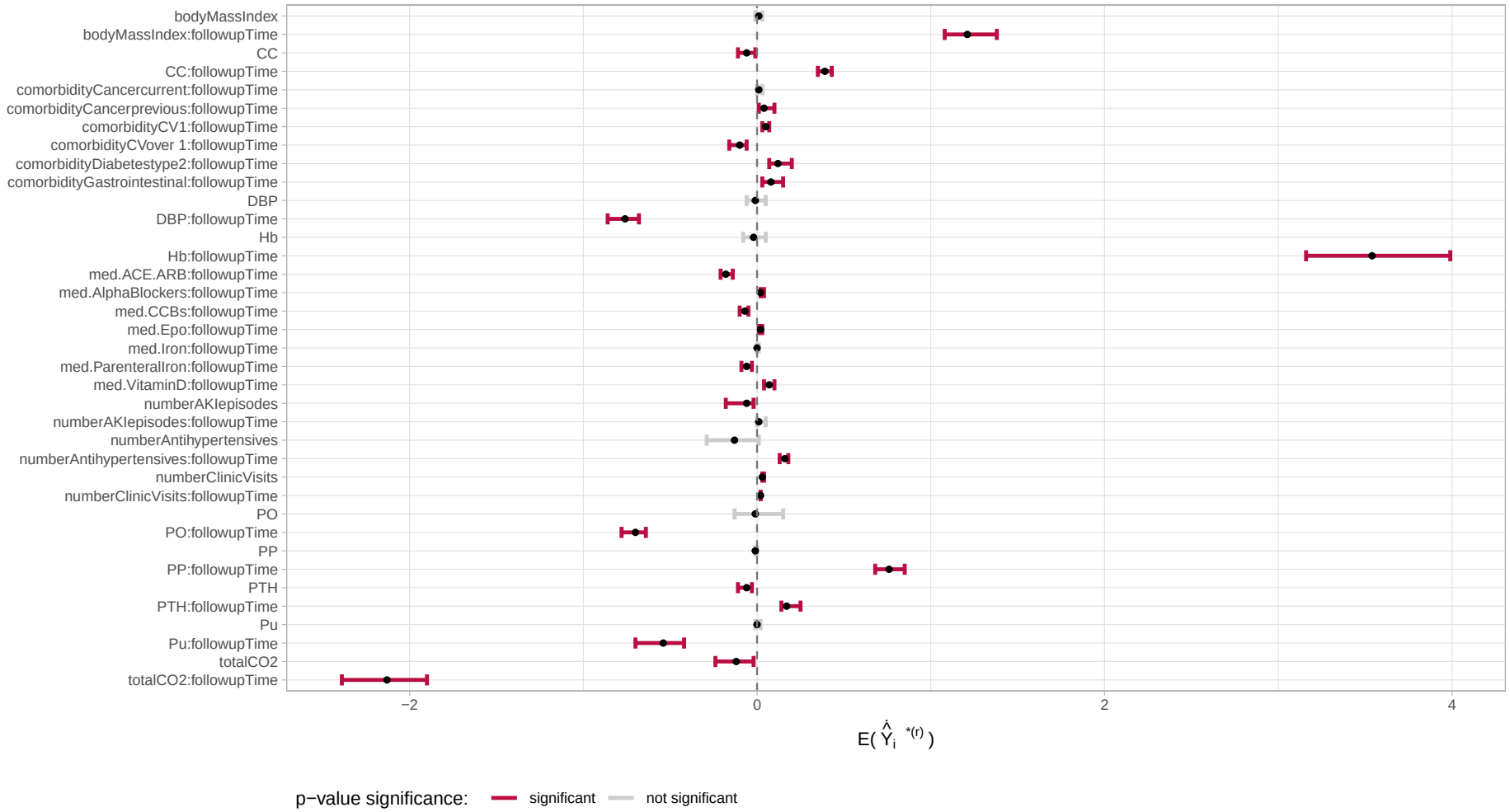

Figure 44: Rate estimates with 95% CIs for pyelonephritis

Table 27: Estimated average rate of change over time for disease pyelonephritis

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-1.04	(-1.46,-0.66)
biochemical		
CC	-0.06	(-0.11,-0.01)
CC:followupTime	0.39	(0.35,0.43)
Hb	-0.02	(-0.08,0.05)
Hb:followupTime	3.54	(3.16,3.99)
PO	-0.01	(-0.13,0.15)
PO:followupTime	-0.70	(-0.78,-0.64)
PTH	-0.06	(-0.11,-0.03)
PTH:followupTime	0.17	(0.14,0.25)
Pu	0.00	(-0.01,0.02)
Pu:followupTime	-0.54	(-0.70,-0.42)
totalCO2	-0.12	(-0.24,-0.02)
totalCO2:followupTime	-2.13	(-2.39,-1.90)
catagorical		
comorbidityCancercurrent:followupTime	0.01	(0.00,0.03)
comorbidityCancerprevious:followupTime	0.04	(0.01,0.10)
comorbidityCV1:followupTime	0.05	(0.03,0.07)
comorbidityCVover 1:followupTime	-0.10	(-0.16,-0.06)
comorbidityDiabetestype2:followupTime	0.12	(0.07,0.20)
comorbidityGastrointestinal:followupTime	0.08	(0.03,0.15)
med.ACE.ARB:followupTime	-0.18	(-0.21,-0.14)
med.AlphaBlockers:followupTime	0.02	(0.02,0.04)
med.CCBs:followupTime	-0.07	(-0.10,-0.05)
med.Epo:followupTime	0.02	(0.01,0.03)
med.Iron:followupTime	0.00	(0.00,0.01)
med.ParenteralIron:followupTime	-0.06	(-0.09,-0.03)
med.VitaminD:followupTime	0.07	(0.04,0.10)
general		
bodyMassIndex	0.01	(-0.01,0.03)
bodyMassIndex:followupTime	1.21	(1.08,1.38)
DBP	-0.01	(-0.06,0.05)
DBP:followupTime	-0.76	(-0.86,-0.68)
numberAKIepisodes	-0.06	(-0.18,-0.02)
numberAKIepisodes:followupTime	0.01	(0.00,0.05)
numberAntihypertensives	-0.13	(-0.29,0.01)
numberAntihypertensives:followupTime	0.16	(0.13,0.18)
numberClinicVisits	0.03	(0.03,0.04)
numberClinicVisits:followupTime	0.02	(0.02,0.02)
PP	-0.01	(-0.01,0.00)
PP:followupTime	0.76	(0.68,0.85)

Table 27: Estimated average rate of change over time for disease pyelonephritis *(continued)*

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.8 Renovascular

On average across the population the dominant terms:

- DBP:followupTime, Hb:followupTime and PO:followupTime contribute to a less rapid decline in kidney function.
- numberAntihypertensives:followupTime and totalCholesterol:followupTime contribute to a more rapid decline in kidney function.

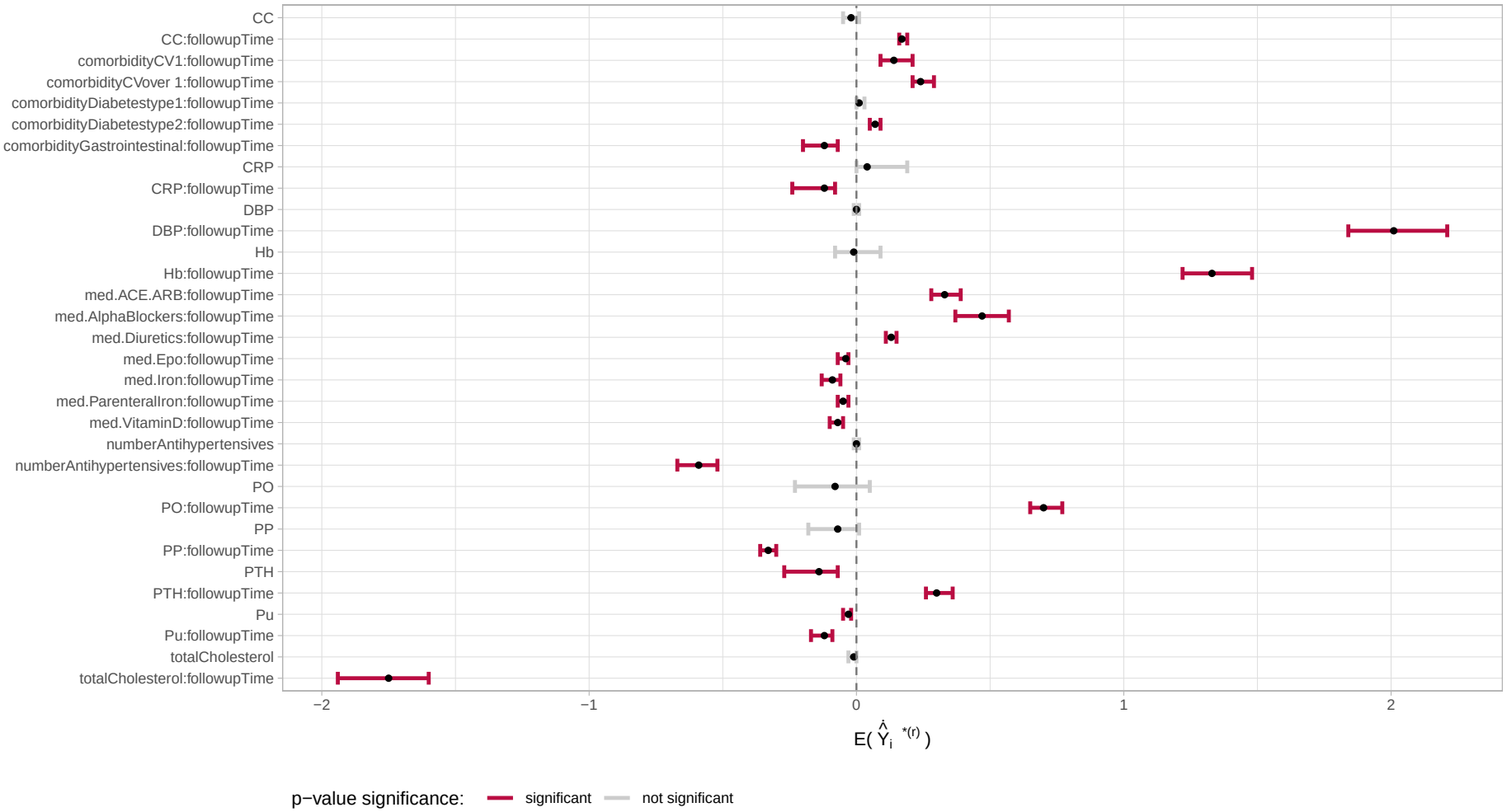

Figure 45: Rate estimates with 95% CIs for renovascular

Table 28: Estimated average rate of change over time for disease renovascular disease

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-0.33	(-0.60,-0.08)
biochemical		
CC	-0.02	(-0.05,0.01)
CC:followupTime	0.17	(0.16,0.19)
CRP	0.04	(0.00,0.19)
CRP:followupTime	-0.12	(-0.24,-0.08)
Hb	-0.01	(-0.08,0.09)
Hb:followupTime	1.33	(1.22,1.48)
PO	-0.08	(-0.23,0.05)
PO:followupTime	0.70	(0.65,0.77)
PTH	-0.14	(-0.27,-0.07)
PTH:followupTime	0.30	(0.26,0.36)
Pu	-0.03	(-0.05,-0.02)
Pu:followupTime	-0.12	(-0.17,-0.09)
totalCholesterol	-0.01	(-0.03,0.00)
totalCholesterol:followupTime	-1.75	(-1.94,-1.60)
catagorical		
comorbidityCV1:followupTime	0.14	(0.09,0.21)
comorbidityCVover 1:followupTime	0.24	(0.21,0.29)
comorbidityDiabetestype1:followupTime	0.01	(0.00,0.03)
comorbidityDiabetestype2:followupTime	0.07	(0.05,0.09)
comorbidityGastrointestinal:followupTime	-0.12	(-0.20,-0.07)
med.ACE.ARB:followupTime	0.33	(0.28,0.39)
med.AlphaBlockers:followupTime	0.47	(0.37,0.57)
med.Diuretics:followupTime	0.13	(0.11,0.15)
med.Epo:followupTime	-0.04	(-0.07,-0.03)
med.Iron:followupTime	-0.09	(-0.13,-0.06)
med.ParenteralIron:followupTime	-0.05	(-0.07,-0.03)
med.VitaminD:followupTime	-0.07	(-0.10,-0.05)
general		
DBP	0.00	(-0.01,0.01)
DBP:followupTime	2.01	(1.84,2.21)
numberAntihypertensives	0.00	(-0.01,0.01)
numberAntihypertensives:followupTime	-0.59	(-0.67,-0.52)
PP	-0.07	(-0.18,0.01)
PP:followupTime	-0.33	(-0.36,-0.30)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.9 Unknown disease

On average across the population the dominant terms:

- Hb:followupTime, numberAntihypertensives:followupTime, PTH:followupTime and totalCholesterol:followupTime contribute to a less rapid decline in kidney function.
- CC:followupTime, DBP:followupTime, PO:followupTime and totalCO2:followupTime contribute to a more rapid decline in kidney function.

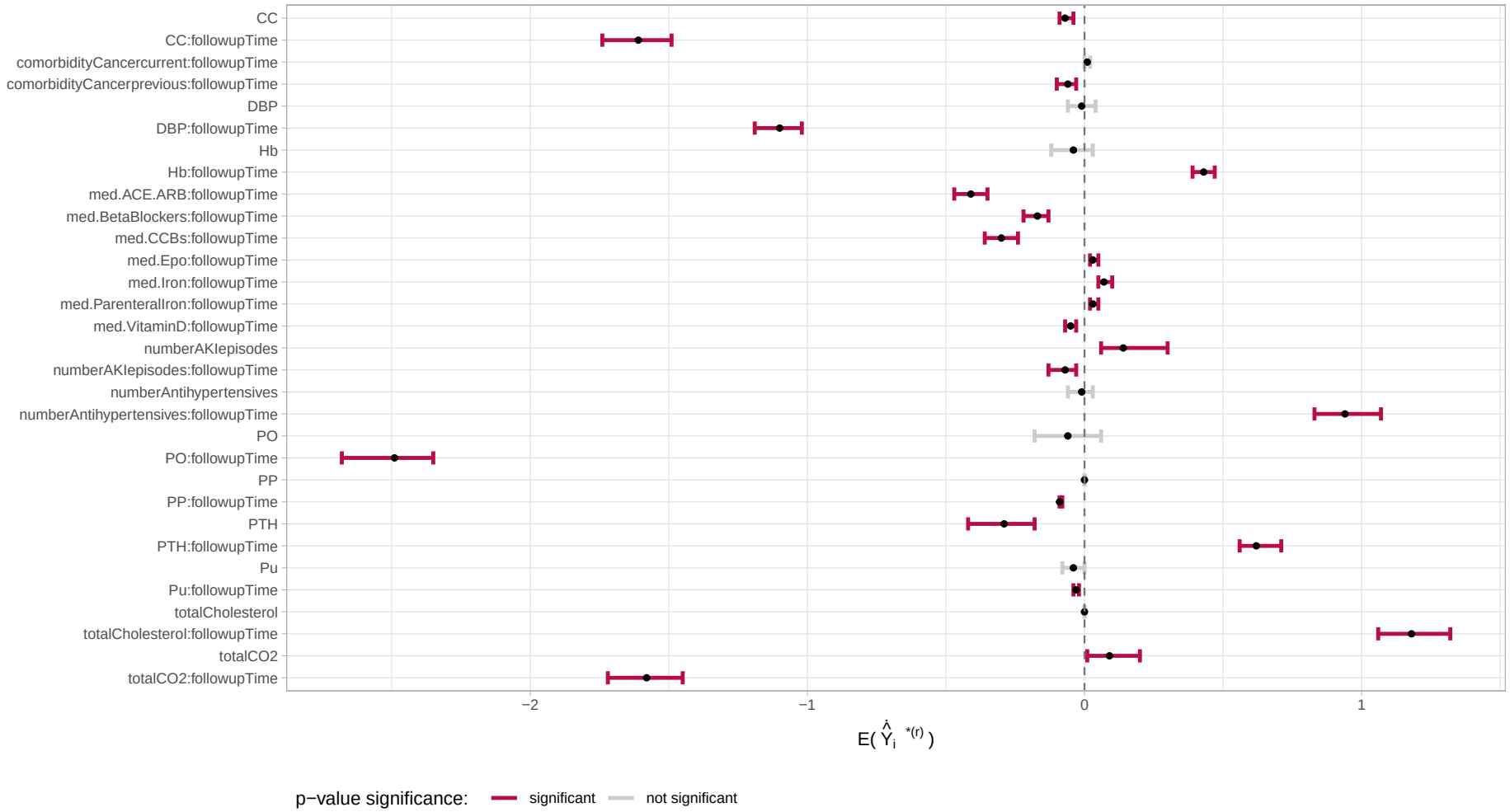

Figure 46: Rate estimates with 95% CIs for disease unknown

Table 29: Estimated average rate of change over time for disease unknown

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-0.42	(-0.70,-0.11)
biochemical		
CC	-0.07	(-0.09,-0.04)
CC:followupTime	-1.61	(-1.74,-1.49)
Hb	-0.04	(-0.12,0.03)
Hb:followupTime	0.43	(0.39,0.47)
PO	-0.06	(-0.18,0.06)
PO:followupTime	-2.49	(-2.68,-2.35)
PTH	-0.29	(-0.42,-0.18)
PTH:followupTime	0.62	(0.56,0.71)
Pu	-0.04	(-0.08,0.00)
Pu:followupTime	-0.03	(-0.04,-0.02)
totalCholesterol	0.00	(0.00,0.00)
totalCholesterol:followupTime	1.18	(1.06,1.32)
totalCO2	0.09	(0.01,0.20)
totalCO2:followupTime	-1.58	(-1.72,-1.45)
catagorical		
comorbidityCancercurrent:followupTime	0.01	(0.00,0.02)
comorbidityCancerprevious:followupTime	-0.06	(-0.10,-0.03)
med.ACE.ARB:followupTime	-0.41	(-0.47,-0.35)
med.BetaBlockers:followupTime	-0.17	(-0.22,-0.13)
med.CCBs:followupTime	-0.30	(-0.36,-0.24)
med.Epo:followupTime	0.03	(0.02,0.05)
med.Iron:followupTime	0.07	(0.05,0.10)
med.ParenteralIron:followupTime	0.03	(0.02,0.05)
med.VitaminD:followupTime	-0.05	(-0.07,-0.03)
general		
DBP	-0.01	(-0.06,0.04)
DBP:followupTime	-1.10	(-1.19,-1.02)
numberAKIepisodes	0.14	(0.06,0.30)
numberAKIepisodes:followupTime	-0.07	(-0.13,-0.03)
numberAntihypertensives	-0.01	(-0.06,0.03)
numberAntihypertensives:followupTime	0.94	(0.83,1.07)
PP	0.00	(0.00,0.00)
PP:followupTime	-0.09	(-0.09,-0.08)

Note:

$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.10 Single model all diseases

On average across the population the dominant terms:

- DBP:followupTime, Hb:followupTime, med.Other:followupTime and PTH:followupTime contribute to a less rapid decline in kidney function.
- CC:followupTime, med.ACE.ARB:followupTime, PO, PO:followupTime, PTH, Pu:followupTime, totalCholesterol:followupTime and totalCO2:followupTime contribute to a more rapid decline in kidney function.

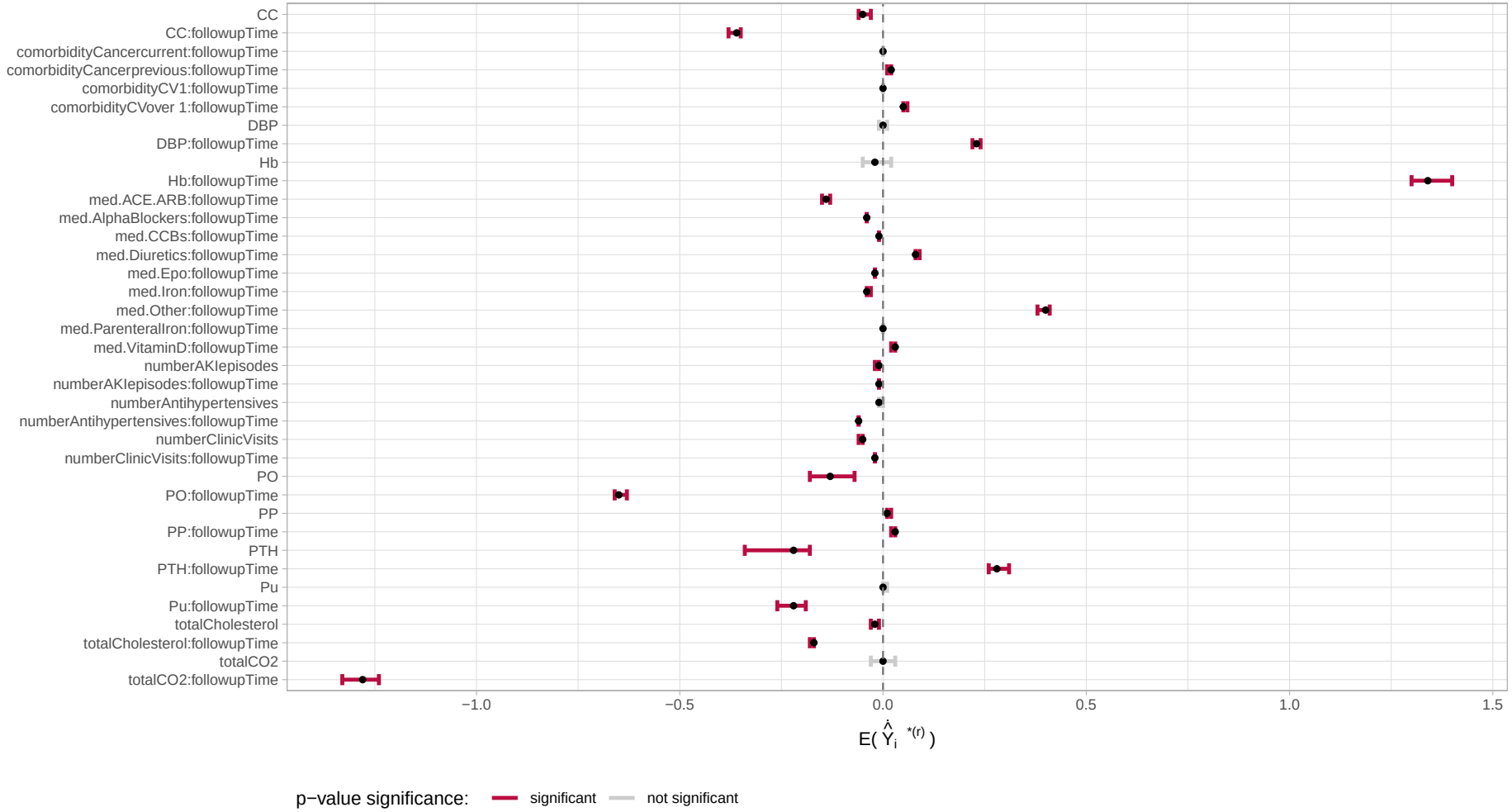

Figure 47: Rate estimates with 95% CIs for single model all diseases

Table 30: Estimated average rate of change over time for single model all diseases

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
overall trend		
outcome variable	-1.08	(-1.20,-0.97)
biochemical		
CC	-0.05	(-0.06,-0.03)
CC:followupTime	-0.36	(-0.38,-0.35)
Hb	-0.02	(-0.05,0.02)
Hb:followupTime	1.34	(1.30,1.40)
PO	-0.13	(-0.18,-0.07)
PO:followupTime	-0.65	(-0.66,-0.63)
PTH	-0.22	(-0.34,-0.18)
PTH:followupTime	0.28	(0.26,0.31)
Pu	0.00	(0.00,0.01)
Pu:followupTime	-0.22	(-0.26,-0.19)
totalCholesterol	-0.02	(-0.03,-0.01)
totalCholesterol:followupTime	-0.17	(-0.18,-0.17)
totalCO2	0.00	(-0.03,0.03)
totalCO2:followupTime	-1.28	(-1.33,-1.24)
catagorical		
comorbidityCancercurrent:followupTime	0.00	(0.00,0.00)
comorbidityCancerprevious:followupTime	0.02	(0.01,0.02)
comorbidityCV1:followupTime	0.00	(0.00,0.00)
comorbidityCVover 1:followupTime	0.05	(0.05,0.06)
med.ACE.ARB:followupTime	-0.14	(-0.15,-0.13)
med.AlphaBlockers:followupTime	-0.04	(-0.04,-0.04)
med.CCBs:followupTime	-0.01	(-0.01,-0.01)
med.Diuretics:followupTime	0.08	(0.08,0.09)
med.Epo:followupTime	-0.02	(-0.02,-0.02)
med.Iron:followupTime	-0.04	(-0.04,-0.03)
med.Other:followupTime	0.40	(0.38,0.41)
med.ParenteralIron:followupTime	0.00	(0.00,0.00)
med.VitaminD:followupTime	0.03	(0.02,0.03)
general		
DBP	0.00	(-0.01,0.01)
DBP:followupTime	0.23	(0.22,0.24)
numberAKIepisodes	-0.01	(-0.02,-0.01)
numberAKIepisodes:followupTime	-0.01	(-0.01,-0.01)
numberAntihypertensives	-0.01	(-0.01,0.00)
numberAntihypertensives:followupTime	-0.06	(-0.06,-0.06)
numberClinicVisits	-0.05	(-0.06,-0.05)
numberClinicVisits:followupTime	-0.02	(-0.02,-0.02)
PP	0.01	(0.01,0.02)

Table 30: Estimated average rate of change over time for single model all diseases *(continued)*

parameter	$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$	CI
PP:followupTime	0.03	(0.02,0.03)

Note:
$\mathbb{E}(\dot{\hat{Y}}_i^{*(r)})$ has units mL/min/1.73m^2/year

7.4.11 Summary

The terms which most frequently occurred overall the disease categories were:

- Hb:followupTime contributing to a less rapid decline in kidney function.
- CC:followupTime, PO:followupTime and totalCO2:followupTime contributing to a more rapid decline in kidney function.

7.5 Counterintuitive results

There are a few clinically counterintuitive results reported above, two of the most commonly occurring relate to PP and DBP. Given the variety and quantity of blood pressure moderating drugs that most patients are taking it is difficult, and beyond the scope of this current work, to draw any conclusions regarding these results. The other frequently occurring unexpected result relates to the interaction term for PTH with follow-up time, we focus on this below.

As reported in Sections 7.2 and 7.3 we consistently found the counterintuitive result that the interaction of PTH with follow-up time is associated with a slower progression of kidney disease, hence it is associated with a shallower slope in eGFR over time. This result is, for example, observed in Table 11. In some LME models the interaction term for PTH with follow-up time (denoted as PTH:followupTime) has a regression parameter with a positive sign and is also statistically significant at the 0.05 level. We seek to determine if this effect is an artefact of the LME model or a real effect present in the SKS data. Throughout this analysis we use the 'single model all diseases' data.

We begin by exploring the data. We compute the average time derivative per patient of a given variable by applying the techniques described in Section 4.2. That is for a given patient and variable with values at discrete time points, we compute the derivative of the spline and then find its expected value to obtain the average slope over time. This allows us to generate Figures 48 and 49 both of which show the average quantities per patient. Figure 48 shows that steeper slopes in log(eGFR) are associated with higher levels of PTH; correlation -0.35. In Figure 49 we observe that steeper slopes in log(eGFR) are associated with steeper slopes in PTH; correlation

-0.43. The results in these figures are clinically plausible, but inconsistent with the aforementioned counterintuitive LME model results relating to the interaction term for PTH with follow-up time. We therefore surmise that this counterintuitive result is an artefact of the LME model and not the SKS data.

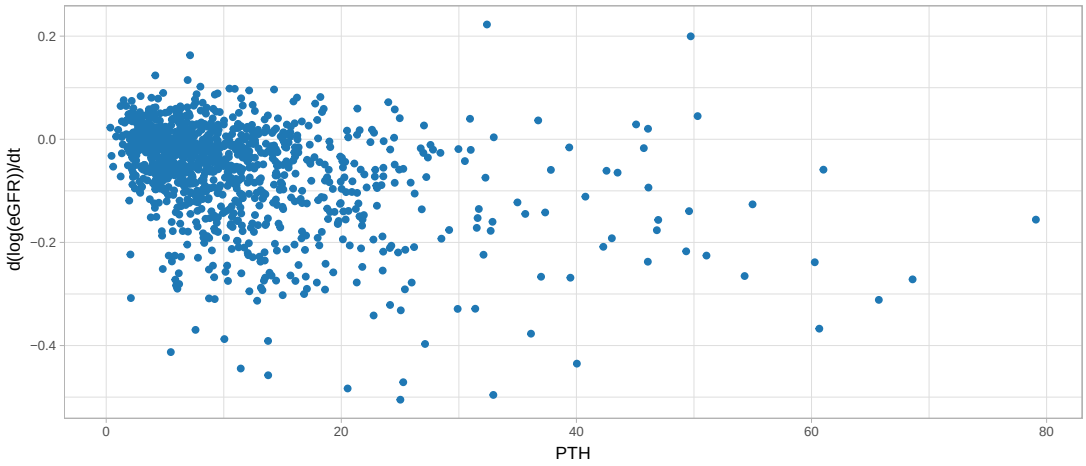

Figure 48: Average time derivative of log(eGFR) per patient versus average PTH per patient

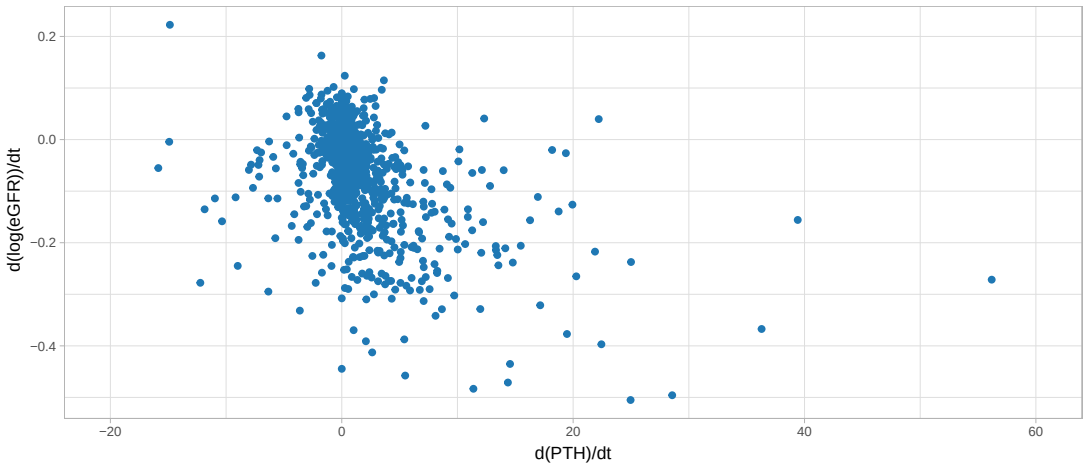

Figure 49: Average time derivative of log(eGFR) per patient versus average time derivative of PTH per patient

We now seek to identify the reason why our LME model may be giving a clinically counterintuitive result for the interaction term relating to PTH and follow-up time. We modify our intercept-and-slope LME model such that the fixed effects are reduced to the following terms $\beta_0 + \beta_1 X_{PTH} + \beta_2 X_{PTH} t + \beta_3 t$; note that t is follow-up time and the outcome variable is log(eGFR). With this reduced model we again find the counterintuitive result that the regression coefficient of the interaction term, β_2, is positive and statistically significant at the 0.01 level. If we use this reduced model but instead use a subset of the data for which the average rate of change of eGFR is negative for each patient, we find that the regression coefficient β_2 is positive but is no longer

statistically significant at the 0.05 level. When the PTH:followupTime regression coefficient is not statistically significant we interpret this to mean that PTH does not markedly influence the progression of disease, that is at population level PTH does not significantly increase or decrease the slope of eGFR.

We conclude, taking into account the reduced LME model using only negative eGFR slopes, that the difference between the PTH:followupTime parameter estimate versus the eGFR slopes in Figure 48 is due to the former being a population level estimate and the latter being an individual level estimate. In the former case our LME model is estimating the effect of PTH on the *average slope* of log(eGFR) given the population, but this *average slope* may not actually exist for any individual. In the latter case we are considering the average effect of PTH on individual log(eGFR) slopes, this is what the clinicians are interested in and what informs their clinical intuition. In summary, the so-called 'counterintuitive' result regarding PTH:followupTime is a consequence of the way our model is summarising the data at population level.

7.6 Correlation between baseline eGFR and its rate of change

In 2013, using the SKS data, Hoefield (85) reported a link between the rate of change of eGFR and baseline eGFR, in particular patients with a higher baseline eGFR had on average a faster decline in kidney function. Disease categories considered were PKD, diabetic nephropathy and glomerulonephritis. In the Results section of (85) Hoefield states, *"Patients with stage 3a CKD at inception into the cohort were associated with more rapid median rates of decline in renal function at -2.06 ml/min/year compared with -1.24, -1.15 and -0.93 ml/min/year in those with CKD stages 3b, 4 and 5, respectively. ... Estimated average decline in eGFR was between 0.8 and 1.6 ml/min/year slower in those patients with eGFR <45 ml/min compared to those with eGFR >45 ml/min at baseline."* However, as discussed in the following, this appears to be inconsistent with our analysis. Note that 45 ml/min equates to 3.8 on the log(eGFR) scale.

By using the techniques described in Section 4.2 we calculated the average time derivative of log(eGFR) per patient, i.e. $\mathbb{E}(\dot{\hat{Y}}_i)$. That is for given values at discrete time points we computed the derivative of the spline and then found its expected value to obtain the average slope over time of log(eGFR). For each patient in each disease group Figure 50 displays the average slope in log(eGFR) against baseline log(eGFR). This figure indicates that for most disease categories patients with lower baseline eGFR tend to have faster rates of decline in kidney function. For patients within each kidney disease category we computed the correlation between baseline log(eGFR) and average slope in log(eGFR). Excluding PKD we found that no disease had a correlation greater than 0.27. In contrast PKD had a relatively high correlation of 0.47. It follows that PKD patients with the steepest negative slopes and fastest rates of decline on average enter the study with lower log(eGFR) values; this correlation is directionally different from Hoefield (85). Similarly, as shown in Figure 50, other diseases in our study also have weak correlations

which are also directionally different from Hoefield (85). We therefore conclude that our results appear inconsistent with Hoefield (85). Future work could investigate the exact details of the data and model used by Hoefield (85) and thereby aim to determine the source of this apparent contradiction.

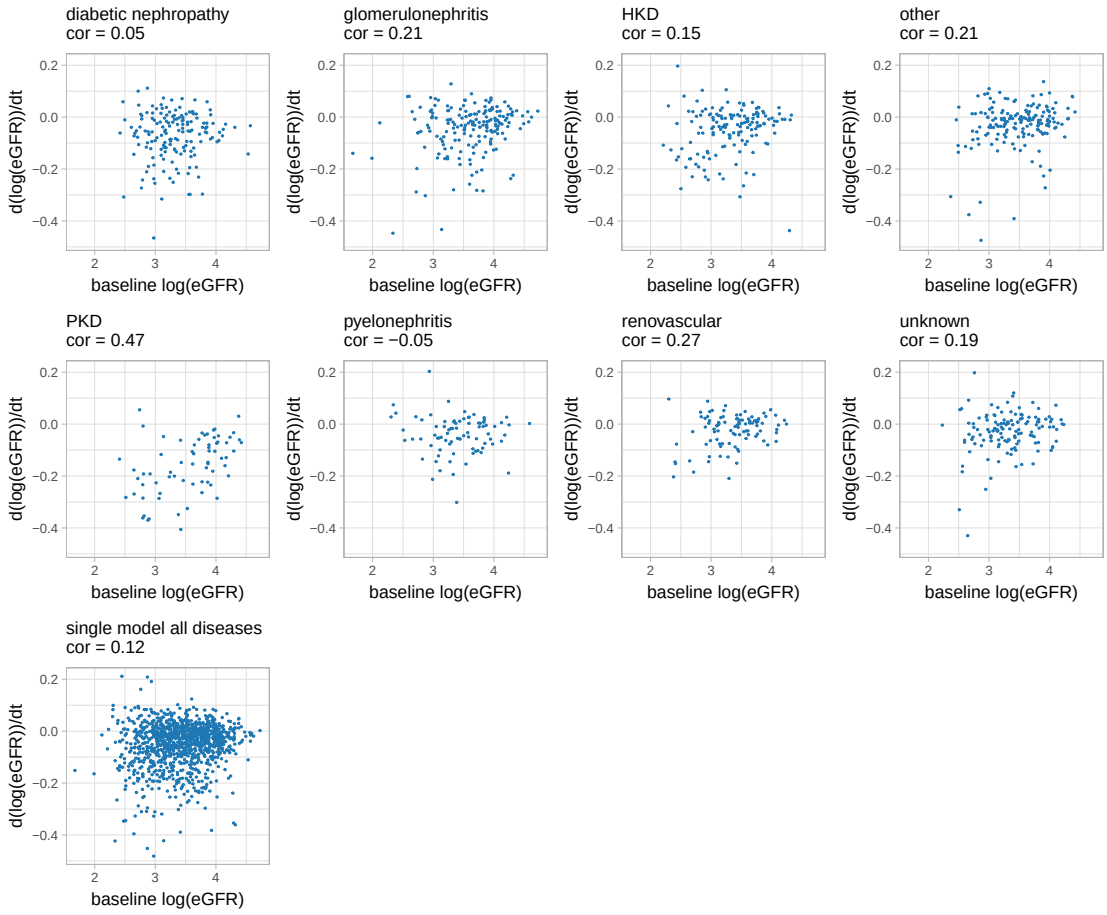

Figure 50: Average slope in log(eGFR) per patient versus baseline log(eGFR)

8 Discussion

8.1 Summary of main results

The aim of this research was to identify key risk factors associated with the progression of kidney disease both across and within eight primary kidney diseases. We used data collected from more than 3000 secondary care patients who had moderate to severe chronic kidney disease and were followed up until they reached an end point of dialysis, kidney transplant or death. Potential risk factors recorded at each annual follow-up appointment included comorbidities, medications, lifestyle factors, socio-demographic information and biochemical marker measurements.

To assess the importance of each of these potential risk factors for each disease we used standard longitudinal modelling techniques, specifically the LME model. We employed commonly used techniques, including bootstrapping, to access model fit and select the pertinent regression parameters for each model. Our main interest was in the fixed effects regression parameters which were used to identify key population level risk factors for each disease category.

First, given our LME for each disease category, we considered the population average level of eGFR: in our models these were the non-interaction terms. We found the risk factors for lower levels of eGFR included biochemical markers and medications, in contrast lifestyle parameters and physical attributes were less important. From a biological perspective we expect the biochemical markers to reflect worsening kidney function, this is consistent with our findings. It may be of future interest to examine the biochemical markers more closely as the differences we observed between diseases could be of clinical value in terms of potential treatment options. Medications play an important role, most notably ACE inhibitors and/or ARBs result in higher levels of eGFR for diabetic nephropathy and glomerulonephritis but not in the other diseases. Also of clinical interest is the role of anaemia management, this enters into our results through iron and EPO medications which we have shown to be strongly associated with diabetic nephropathy. In our results lifestyle parameters did not play a significant role, this could be because these risk factors contained so much missing data that we found it necessary to reduce them to baseline variables (e.g. smoking status, occupation, alcohol intake). As expected, baseline age was a clear risk factor. Body mass index was only strongly associated with diabetic nephropathy, this association is unsurprising in light of the fact that being overweight is a risk factor for type 2 diabetes. Interestingly, although not significant, the model did identify sex as associated with pyelonephritis. This association is clinically plausible given females are more susceptible to urinary track infections which can contribute to pyelonephritis.

Secondly, given our LME for each disease category, we consider the rate of progression of eGFR over time. In our models these are the interaction terms with time, i.e. $explanatoryVariable : followupTime$. These temporal effects are harder to interpret clinically as they relate to a

population trajectory which may never be observed in clinical practice where the focus is on individuals. This population level effect has in part led to results which go against clinical intuition, e.g. high levels of PTH being associated with less rapid decline in kidney disease. However our model did identify that rapid progression of kidney disease is associated with biochemical markers including PO, PTH and total CO2: this is biologically plausible. In general we found that medications and comorbidities were not key in rapid disease progression. In the future more work is needed to consider a wider range of statistical methods which could lead to clearer identification of risk factors relating to the progression of kidney disease over time. Perhaps a more nuanced approach could better identify risk factors such as comorbidities and/or medications.

Thirdly, we used a more novel approach towards understanding disease progression by considering time derivative of the fitted LME model. We found, given all disease categories, that PKD had the most rapid progression of kidney disease, with a loss in eGFR of 3.5 mL/min/1.73m^2/year whereas the rest of the categories show a loss of around 0.5-1.5 mL/min/1.73m^2/year: these results are consistent with previous work. In addition we also reported the breakdown of rates for each risk factor: these results were consistent with our aforementioned results. This is unsurprising given they have the same model at their foundation. Given each continuous variable it may have been more informative to combine the non-interaction and interaction term into a single term, this would then describe all time variability of the given risk factor in single quantity. This approach might better describe the relative importance of the risk factors and could lead to a resolution of the counterintuitive results mentioned in the previous paragraph. Further research is needed into the best approaches for determining risk factors in relation to rates of change over time of an outcome of interest.

Many of the risk factors we identified match clinical intuition and thereby confirm what is already known in clinical practice. However in some instances our results point towards a need for further clinical studies, most notably the common clinical practice of prescribing ACE inhibitors and ARBs regardless of the kidney disease type. Our identification of key risk factors relating to kidney disease progression has implications for the monitoring and treatment of future chronic kidney disease patients. Below in Sections 8.4 and 8.5 we discuss these in more detail, specifically potential implications in terms of mental health, socio-economic factors, medications and personalised healthcare.

8.2 SKS Data

8.2.1 Strengths

Since 2002 the ongoing SKS study has recruited over 3000 secondary care patients who are followed-up annually. SKS has well defined end-points which in part ensures patients have comparable stages of kidney disease; patients are removed from the study once they commence

RRT or die. In addition biochemical data is collected outside follow-up appointments, during routine clinic visits and during acute episodes of illness e.g. AKI. This is a very rich dataset containing, after cleaning and before imputation, 1,103,163 distinct units of information. SKS is one of the largest and longest ongoing studies of its kind in the world. There is a similar ongoing study, The Chronic Renal Insufficiency Cohort (CRIC), which since 2001 has recruited about 5500 adult CKD patients from 11 clinical sites across the United States, this study also has the core aims of investigating risk factors for CKD progression and links to cardiovascular disease (86,87). There is a significant ethnic difference between SKS and CRIC in that by design CRIC has 40% African American and 10% Latino/Hispanic or Asian/Pacific Islander. Both studies capture a broad range of potential risk factors for each patient, these include biochemical, comorbidity, medication, lifestyle and socio-economic demographic data.

8.2.2 Limitations and weaknesses

There is clearly a limit on that which can reasonably be recorded at a follow-up appointment by a clinician whose first priority is to patient care rather than data collection. This trade-off has resulted in considerable incomplete records in the data, in total about 71% of all follow-up appointment data has at least one field missing; mostly because it was not recorded but sometimes because it was incorrect so was deleted by our 'data cleaning' procedures. In addition the biochemical data collected from each patient at their follow-up appointment typically takes several days to be processed and is recorded in a separate database. This makes matching the biochemical data with the SKS data non-trivial, if no match is found and imputation is not possible the whole follow-up appointment record is unusable when modelling. If creatinine is not measured at a follow-up then all the data from that appointment will be discarded as we do not impute the model outcome variable.

The significant quantity of missing data resulted in us resorting to imputation methods to gain statistical power in our models. This meant we had to make several pragmatic decisions as follows. If a patient is recorded as having a comorbidity we assume they have the condition for all future time, which is a reasonable assumption for all chronic conditions. We relied on the first instance of a comorbidity being recorded correctly; if a given patient was mistakenly recorded as having a particular comorbidity we then propagate this error through all their subsequent follow-ups since there is no mechanism by which we can determine the recording mistake. At a given follow-up, if at least one medication is recorded then we assign the patient as either taking, or not taking, medications in all of the aforementioned medication categories. We therefore assume all medications have been correctly recorded in the SKS data at the follow-up. Only at follow-up appointments where no medications are recorded do we impute this data.

Where possible continuous variables are imputed using the Kalman method as described in Section 2.7. However this does not allow upper and lower bounds on the values it returns so

potentially it could return negative values. This would lead to erroneous imputed values since all our continuous variables are take positive values. Our pragmatic solution was first to transform each field onto the log scale, secondly perform imputation by the Kalman method, and finally transform back to the original scale. We expect this to cause some minor dependencies since a log transformation will result in an approximately linear imputation on the log scale instead of on the original scale. However if the measured values are falling or rising according to a power law on their original scale, as many biological systems/markers do, they will be linear on the log scale in which case linear imputation is ideal.

8.2.3 Recommendations

We used imputation to replace missing values, but this was performed on each subject separately where each field for which we imputed values was treated as a timeseries. All subjects and fields were therefore treated as if they were mutually independent and all values once imputed were treated as if they were real measured values with no acknowledgement of uncertainty; clearly this could lead to over-confidence in the model results and future work should consider addressing this. It would be of interest to investigate more sophisticated imputation methods. Multiple Imputation by Chained Equations (MICE), e.g. see (58,88), is a very popular method which uses a series of regression models where each variable with missing values is modelled conditional on the other variables within the data. As such each variable can be modelled with respect to its distribution; e.g. continuous variables are modelled with linear regression and binary variables with logistic regression. Assuming we believe all values are missing at random then a timeseries extension of MICE should improve our imputation, but to the author's knowledge no such method exists. Perhaps in the future such a method could be developed. With a view to improving upon our strategy of transforming to a log scale prior to imputation, future work should also consider imputation methods that would assure the imputed (biochemical) values were always positive. For categories, such as medications, where drugs could be prescribed intermittently rather than attempting to impute data it may of interest to investigate if it would be more appropriate to give a medication category three levels 'taking drug', 'not taking drug' and 'unknown if taking drug'. The final level 'unknown if taking drug' may allow for the model to take better account of the missing data.

The original raw SKS dataset was far from clean as detailed in Appendix A.1. As a result considerable time and effort was spent cleaning it before any analysis could commence. We recommend, for example, recording all measured values of a given variable in the same units. Furthermore many issues with poorly recorded or missing data could relatively easily be overcome by incorporating user-friendly front-end software on the SKS Microsoft Access database. This front-end software could be used by the clinician or nurse to enter the data into the database; Figure 51 shows an illustrative example.

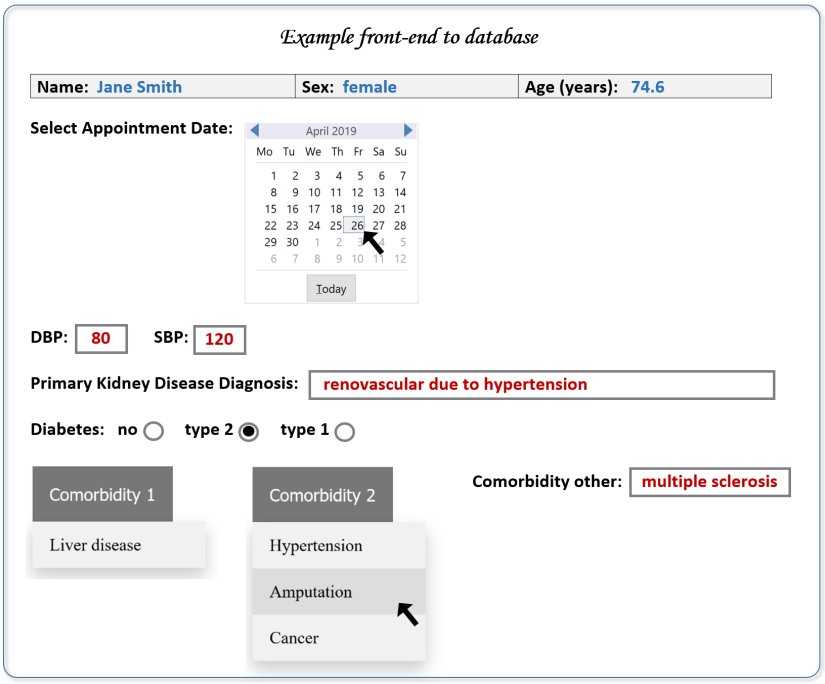

Figure 51: Illustration of front-end software for entering data to database

Benefits of front-end software include:

- when appropriate, automatically update fields using data from previous follow-ups (e.g. chronic comorbities, such as amputation) so that once it is recorded it is automatically recorded at all future appointments unless the user takes action to remove it (e.g. if amputation was erroneously recorded)
- appointment dates would be accurate and use a consistent format i.e. free from reverse ordering year/month/day
- units of measurements would be consistent e.g. height in metres and weight in kilograms
- database table names and table column headings would by default use consistent naming conventions
- the front-end software could automatically run checks to make sure certain types of data were realistic e.g. dates of birth, DBP is less than SBP, remove erroneous characters
- where possible drop-down menus would contain lists (e.g. of comorbidities, medications) to ensure consistent data and avoid erroneous character entries and misspellings
- free text boxes could be used for additional information e.g. less common medications or rarer comorbidities
- the medications the patients were taking at the last follow-up could be displayed on the screen, the accuracy of the patient self-reporting or a clinician reading other medical notes could perhaps be improved if the medication changes since the last follow-up were more obvious.

However such a front-end would not be without its drawbacks. The front-end would need

maintaining, for example it may need updating if the underlying structure of the database changes e.g. a new field is added. Moreover if the front-end constrains the range of values too much this could potentially make it impossible to record unusual/unexpected values. Such outliers could represent vital information from both a medical and/or research perspective.

Front-end software can be added to any Microsoft Access Database, in the past the author of this thesis has programmed such software using SQL embedded in Visual Basic. Such software could not guarantee clean data but it could hope to reduce both missing values and recording mistakes. Overall, front-end software could dramatically improve the data quality, reduce the need to impute data, substantially decrease the quantity of intricate data cleaning code, and crucially improve the statistical power of the models.

8.3 Statistical model

8.3.1 Strengths

We have shown that our models have good fits for all disease categories and where known that our parameter estimates are clinically plausible. These claims are in part supported by both the literature and expertise of the SKS clinicians. We used multiple bootstrap samples during the model selection procedure to help guard against overfitting to our data. This also allowed us to take into account parameter uncertainty when selecting the final model for each disease. We were able to utilise the LME model which is a long-standing standard framework frequently used for longitudinal analysis. This allowed us to use established software for fitting our models. Unlike the Generalised Estimating Equation (GEE) framework, another standard approach for estimating parameters in longitudinal linear regression models, the LME can accommodate missing longitudinal data (e.g. missing annual follow-ups) and differing lengths of time in the study; these characteristics are abundant in our data.

8.3.2 Limitations and weaknesses

For model selection we used a stepwise regression procedure with AIC as there were a large number of potential explanatory variables and no underlying theory on which to base model selection beyond some advice from the clinicians. To make selecting explanatory variables from a large pool more robust to overfitting we repeated the stepwise regression procedure on 100 bootstrap samples for each disease model. We consider this an acceptable number of bootstrap samples. However, although more bootstrap samples would obviously add greater statistical weight to our parameter selection, we were constrained by computation time. The largest dataset was for 'single model all diseases' which took ~100 hours to process with 100 bootstrap samples on a single processor (Intel(R) Core(TM) i7 CPU @ 2.70GHz); we parallel processed by disease

model. The final model for each disease was chosen using parameters which occurred in more than 50% of all bootstrap samples but this percentage was an arbitrary choice. Note that the lengthy computation time meant we only split our data into training and validation data as cross-validation was impractical. If future work aims to check overfitting on a subset(s) of data it should consider using cross-validation techniques.

Fitting the final selected model obtained using stepwise regression has a long history of being criticised in the literature especially when parameter estimates, p-values and confidence intervals are reported without adjusting them to account for the model building process e.g. see (89,90). For example, Harrell (91), states that parameter estimates are biased away from zero, while standard errors and p-values are biased toward zero. In practice there may be no reasonable way of correcting for these problems. We acknowledge the stepwise regression procedure with AIC is imperfect which is why we employed it within a bootstrapping scheme. A deep analysis of uncertainties which accounts for the model building procedure is beyond the scope of this thesis, hence our final selected models are reported without incorporating model selection uncertainty. We also note that we report results without accounting for uncertainty induced by imputing missing values.

Machine learning involves using algorithms which can learn from, and make predictions on, data. In this context it is commonplace to fit a model to training data and assess the relevant aspects of the fit using validation data; for example see (92) for a review relating to the evaluation of regression models. In terms of a regression model, such as an LME, if predictions are the aim then it is reasonable to fit the model to the training data and assess the quality of predictions using validation data. Here our focus is on identifying the regression parameters pertinent to each disease category. However we have a large pool of regression parameters each of which may, or may not, be selected in each of our final disease models; as a result overfitting to our data is a potential hazard. We would like our model to generalise to future SKS data and new datasets, therefore in our context a legitimate concern is overfitting to the current SKS data. To this end we undertook model selection on the training data and then determined if the final model selected also fitted within reason to the validation data. In Section 6.2, we examined parameter estimates, confidence intervals and residuals, given each disease's final model for each dataset (training and validation). This approach is open to criticism, the small sample size of the validation data, as compared to the training data, will always result in wide confidence intervals on parameter estimates. As a result a comparison of parameter estimate confidence intervals between these two datasets is a weak test of overfitting. Although for LME models the maximum likelihood estimates are consistent, the bias in a small sample is potentially large. As a consequence differing estimates between training and validation datasets are a necessary but not sufficient condition for concluding that overfitting has occurred. If one of our models had failed to fit to the validation data then this would have alerted us to the possibility of a potential problem with that model fit and would have prompted further investigations into the selected model. After selecting the model

with our bootstrapping procedure we used training and validation data on the understanding that it offers no more than a final weak test of overfitting. Future work should consider the use of training and validation datasets in greater detail, if the decision is made to use them, then robust evaluation procedures, such as those described by (92), should be implemented.

A further shortcoming of the use of training and validation datasets was that it would have prevented us from creating a model for disease 'obstruction' as there was insufficient data; in total (before splitting) there were 23 patients with a total of 49 follow-up records. It was beyond the scope of this thesis to consider constructing a separate model for this small group of patients.

We acknowledge that the LME model assumes any measurement error in explanatory variables is negligible compared with the within-group errors ϵ_{ij} (see Equation 2), in essence we are treating explanatory variable observations as exact; this is the usual assumption in the literature. Although beyond the scope of this thesis, it would not be unreasonable in the future to consider the extent to which this assumption is reasonable in clinical practice particularly in relation to biochemical measurements and other biological measurements such as blood pressure. If this assumption was shown to be suboptimal then such explanatory variables may be both time-dependent and stochastic; for example see (65) Chapter 12.3 for a framework relating to longitudinal models with stochastic covariates. A way forward would be with a joint modelling framework where each such variable in our existing LME model is replaced by a stochastic process; this could be fitted using a 2-stage process which first fits each explanatory variable model and secondly fits the longitudinal LME model.

We observe a noteworthy quantity of autocorrelation in the residual plots of Figures 8-16. We explored all relevant predefined correlation structures (compound symmetry (CS), first order continuous-time autoregressive (CAR1) and a general correlation matrix with no additional structure) available in the software we used (i.e. R-package nlme (66,67)) but none of these substantially reduced the residual autocorrelation. In the following we refer to two of the models, Model C and Model D, defined in Section 5.4 *Step 4.*; note that thesis results are based on Model C. Both models have identical fixed effect terms along with intercept-and-slope random effect terms. These models differ only in their correlation structure. In Model C the within-group correlation structure $\boldsymbol{C}_i$ is not defined, whereas Model D has correlation $\boldsymbol{C}_i$ defined as the CAR1 model. We use as an exemplar renovascular disease to plot the autocorrelation for Models C and D, see Figure 52. Model D with correlation structure CAR1 is barely an improvement on Model C in which a correlation structure is not defined. Within the scope of this thesis we were unable to better, or more fully, account for the correlation structure inherent in the data.

As previously described, when selecting the model in Section 5.4 we encountered problems specifically relating to difficulties fitting multiple bootstrap samples when using a CAR1 model (i.e. Model D). Future work should consider in far greater detail the within-group correlation structure of the LME model in the context of the data. If the difficulties we encountered can be

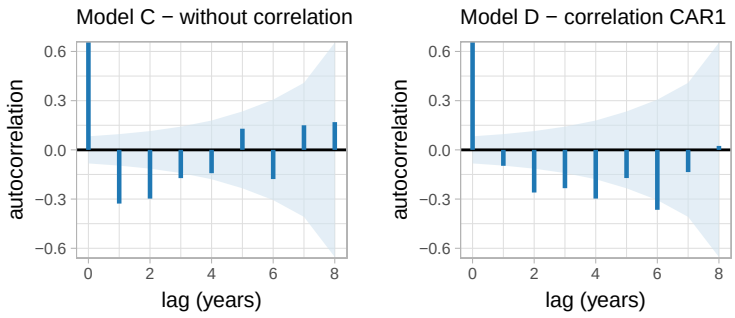

Figure 52: Residual autocorrelation for Models C and D

overcome then likelihood ratio tests, rather than pragmatic decisions, should be used to determine, for each disease separately, the model which best describes the data. Different within-group correlation structures may be optimal for different diseases.

In Section 4.2 we considered the time derivative of the LME model. By focusing on the fitted model we were able to ignore the stochastic nature of the LME model given in Equation 2. It would be of future interest to consider in much greater detail the time derivative of the LME model prior to fitting. This would involve taking into account the stochastic nature of the model which is encoded in the within-group error term; this term may also include an explicit within-group correlation structure. If the time derivative of the LME model can be fitted this may better address the issue of understanding the progression of kidney disease on an individual level and circumvent the population level counterintuitive results discussed in Section 7.5.

The residual plots in Figures 8-16 show that our LME models do not account well for all patients as there are some outliers. This does not cause us concern in relation to determining fixed effects which is our main focus but it would be problematic if we were interested in predictions; we could not make reliable predictions for outliers. Our models assume that random effects and within-group errors follow multivariate normal distributions. However if we replaced these distributions by their corresponding multivariate t-distributions with identical means and variance-covariance matrices then the random effects and within-group errors should be more robust to outliers; heavier/longer tails will better accommodate some, if not all, outliers. If we were interested in predictions this would almost certainly result in more robust inference; e.g. see (93) Chapter 9 for details.

8.3.3 Recommendations

The correlation structure chosen for an LME model will directly affect any predictions made using the model. If in the future our LME models were to be applied to patient-specific predictions of kidney disease progression then it would be necessary to determine an appropriate correlation structure for each LME model presented in this thesis; it may be that different diseases require different correlation structures.

Sometimes the biochemical markers measured at annual follow-up appointments are also measured between these appointments when the patient attends a clinic or is admitted into hospital. At present we are not including these SKS data in our analysis. It may be of interest to investigate ways of utilising this additional information in our models, especially when it is possible to confirm that the measurements did not relate to an acute episode of illness, such as AKI. One way to utilise this extra data would be to use it to aid the imputation of biochemical measurements at follow-up appointments.

In instances where the model could maintain sufficient statistical power it would be of interest to sub-divide the heterogeneous disease categories used in this thesis. In particular the clinicians advised that it would be of future interest to sub-divide the heterogeneous disease category 'other' and also sub-divide glomerulonephritis into about 4 sub-categories. Likewise some of the comorbidity categories could be sub-divided, notably the cardiovascular category is particularly heterogeneous with a mix of chronic and acute conditions. Similarly it may prove informative to include more medication categories; this is a very rich data source within the SKS dataset which in the future could be exploited more fully.

8.4 Implications regarding disease progression

8.4.1 Mental Health

It may be of interest to consider recording each patient's mental heath state despite, to the author's knowledge, a directional causal link from mental to physical health not being clearly identified in the literature. Such a link may be difficult or impossible to identify, given a person's mental and physical health may interact with each other in very complex ways; future scientific advances can hope to precisely identify the biological mechanisms. For example (94) investigates how past physical/mental health influences current physical/mental health (a clear directional link was not found) and (95) discusses how negative emotions could be responsible for diseases whose onset and course may be influenced by the immune system. Scott et. al. (96) uses data from 40,000 adults from 17 countries to assess the relationship of depression and/or anxiety with chronic physical conditions. They conclude that their work points towards: a) mental health disorders leading to physical conditions; or b) the same factors being conducive to both multiple mental health disorders and physical conditions. Poor mental health, e.g. depression or anxiety, may contribute towards a patient making poor lifestyle choices, such as insufficient exercise, poor diet, drinking too much alcohol, etc. In terms of CKD these choices could in turn contribute to an increase in the patient's risk of CKD and/or rapid kidney disease progression. For instance (97) studies how anxiety and depression are associated with unhealthy lifestyle in patients at risk of cardiovascular disease. To the author's knowledge there is no equivalent study for CKD.

8.4.2 Socio-economic factors

In our models we did not find a strong connection with socio-economic risk factors, although these are likely to influence patient lifestyle choices, therefore perhaps further consideration could be given to these factors as alluded to in the previous section. For example perhaps for each patient it may be of interest to record their Index of Multiple Deprivation (IMD), lifetime earnings, education level, level of engagement with primary care health services, and so on.

8.4.3 Disease progression with respect to baseline eGFR

In Section 7.6 we reported our finding that on average PKD patients with lowest baseline eGFR also had the fastest rates of decline in kidney function. Ideally these patients need to be referred from primary to secondary care much sooner so as to explore treatment (medication) options which could slow their rapid kidney function decline to end point (RRT or death).

Our observation of steeper eGFR decline correlating with lower baseline values relates to most disease categories which we studied. On a population wide level, as shown in Figure 50, this relation does not apply to every patient. Although it does seem reasonable to expect more rapid decline to be associated with low baseline eGFR, this correlation would appear to be in conflict with the findings of Hoefield (85). Given Hoefield and ourselves both used SKS datasets, future work could compare the differences between our respective models and data. For example such analysis should include: a) evaluation of differences in data cleaning procedure; b) the effect of the updates carried out by the SKS clinicians in 2019, which were applied to our dataset, that related to confirming or reassigning the primary disease category of each patient; and c) the consequences of our study using an extra four years of data.

8.5 Future work

8.5.1 Joint longitudinal and survival modelling

A natural extension of the work presented in this thesis would be to give consideration to survival analysis which accounts for time until a pre-specified event occurs, i.e. time-to-event; such methodology is widely used, for example see textbook by Hosmer (98) and reviews (99,100). In survival analysis the 'risk set' contains patients at risk of experiencing an event, this consists of patients who have been followed-up until a certain time but have not yet experienced the event of interest e.g. death. Survival analysis accounts for the fact that the survival time is censored for patients who do not experience the outcome of interest within the observation period, furthermore it is unknown when, or whether, such patients will experience the event in the future. Consequently censored time-to-events are a type of incomplete data, specifically they occur

when the patient: a) is lost to follow-up; b) experiences another event which makes follow-up impossible or meaningless; c) has not experienced the event of interest within the observation period. Censoring is assumed to be independent and randomly occurring, that is at a given time point patients who are censored are representative of those still at risk; e.g. it is assumed to be a random occurrence if a patient is lost to follow-up due to migrating a significant distance and that such a patient will be at a similar risk of experiencing the event of interest as one who is still in the study.

In the SKS data the time-to-event would be time from baseline to RRT or death. There is significant censoring in this data given 99 patients were recorded as lost to follow-up and a further 699 patients had a gap of 2.5 years or more since their most recent follow-up appointment. Our longitudinal models take no account of these events so they overlook this potentially informative source of censored information. Survival analysis provides an appropriate framework to account for time-to-event data.

Given the SKS data we propose that survival analysis is undertaken with a variant of the Cox model (101,102) that allows for time-varying explanatory variables e.g. see discussions (103), (104), and Hosmer book (98) Chapter 7.3. We note that Asar (105) demonstrates for kidney data, similar to ours, that joint modelling with a Cox proportional hazard regression model and a longitudinal LME model is better than performing separate longitudinal and survival analyses. This is because the joint model makes optimal use of all available data and correctly handles irregularly measured time-varying explanatory variables, thus the joint model achieves unbiased estimates of model parameters (105). Consequently future work could investigate the use of this joint modelling framework to develop, for each kidney disease category, an understanding of how the typical pattern of disease progression is influenced by time-to-event data along with baseline and time-varying explanatory variables. In addition it may also be informative to investigate constructing a joint model that is specifically designed to investigate how treatment effects influence time-to-event outcomes.

In the context of CKD progression it would also be interesting to study the effects of competing risks. These occur when patients could potentially experience one or more events or outcomes which 'compete' with the outcome of interest. The competing risk either modifies the chance that the event of interest occurs, or masks/hinders its observation. Of particular clinical interest in our case are the competing risks of cardiovascular disease, initiation of RRT and/or mortality. However these competing risks are asymmetric in that cardiovascular disease and RRT would precede death, but death automatically censors cardiovascular disease and initiation of RRT. Likewise in the SKS study initiation of RRT automatically censors cardiovascular disease. See Lau (106) for a general discussion of competing risk methods, Noordzij (107) for a discussion of competing risks methods for survival analysis applied to kidney disease and (108) for details regarding joint modelling of longitudinal and survival data in the presence of competing risks.

It would be of interest to consider extending our LME models for longitudinal data to latent class linear mixed models which additionally account for unobserved heterogeneity within a population (patients in a disease category). This heterogeneity is modelled by classifying 'similar' individuals into unobserved sub-groups (latent classes) each of which is less heterogeneous (i.e. more homogeneous). In the longitudinal case each latent class is characterised by its own mean trajectory; e.g. see overview by Berlin in the two part article (109) and (110). Furthermore, it could also be informative to consider a joint modelling framework of a latent class mixed model for longitudinal data with a survival model for time-to-event data. In such a joint model individuals would be characterised by a class-specific linear mixed model (with fixed and random effects) for their longitudinal data along with a class-specific survival model for their time-to-event data. Details of this type of joint modelling where time-to-event data is accounted for using proportional hazard models are given by Proust-Lima (111) and implemented in the R-package LCMM (112). At present we do not know if there are any informative latent classes within our disease models but it would be worth exploring this avenue. For example if a clear indication of latent classes was found in the glomerulonephritis disease category this could suggest that this category, which we know to be heterogeneous, should be split by the clinicians into several sub-disease categories; these sub-categories could then be modelled separately. Alternatively, if we found that a well-defined homogeneous disease category, e.g. PKD, appeared to have two or more latent sub-classes we would consider, in conjunction with the SKS clinicians, the possibility that there were one or more key explanatory variables missing from this disease model.

8.5.2 Personalised healthcare

An area of increasing interest throughout the health sector, including the NHS, is personalised healthcare (113), that is the individualisation of treatment by identifying patients who are most likely to respond positively to a particular treatment regime. This is beneficial to patients as they receive the most appropriate treatment plan and has the added advantage that healthcare providers do not waste funding on treatments that deliver insignificant benefit, no benefit or cause harm to the patient; for example see discussions (114–116). The models in this thesis have the potential to contribute towards personalised healthcare. For example we have shown that ACE-inhibitors and ARBs are particularly beneficial to diabetic nephropathy and glomerulonephritis patients but on average are not beneficial to the remainder of the cohort; the clinicians advise us that they routinely prescribe these drugs across the entire cohort regardless of primary kidney disease. Future work could hope to more accurately identify each patient that would benefit from ACE-inhibitors and/or ARBs.

In the future our models could be used to pave the way towards personalised real-time predictions of kidney disease progression, assuming the following issues can be adequately resolved: outliers; LME model within-group correlation structure; and cleaner recording of patient data in the SKS

database. This thesis uses longitudinal LME models to explore population effects but such models have also been considered for personalised predictions; for example a longitudinal LME model for personalised real-time predictions of primary care CKD patients which uses a few covariates (e.g. baseline age, comorbidity indicators and sex) is proposed by Diggle (117).

However, if personalised prediction is a future aim, it is commonly advised in the literature that joint modelling of longitudinal and time-to-event data improves the predictive capability and so leads to more informative inferences; e.g. see review by Hickey (118) and also Brankovic (119) who focuses on joint models for personalised prognosis in CKD patients. In this context for each kidney disease category we would envisage a joint model consisting of a Cox proportional hazard model and a longitudinal LME model of the kind explored in this thesis. However for personalised prediction models to be of use in clinical practice they will need to be straightforward to interpret by the clinician, the underlying data will need to be sufficiently clean and up-to-date, and the models will need to undergo rigorous validation (120).

We envisage personalised predictions could, using a web page linked to the SKS data, display a graph of the predicted rate of decline of kidney function for a given patient at the time of their follow-up appointment; illustration given in Figure 53. Furthermore an algorithm could be embedded in the web page to predict which treatments would be most likely be beneficial to the patient conditional on the population average for their given primary kidney disease type; e.g. how would the patient's trajectory change if they started taking a β-blocker. Provided the web page was designed to be user-friendly there is no reason why the clinician, and perhaps an interested patient, could not look at the predicted trajectories during the follow-up appointment. Such a system would not replace the expertise of the clinician but it could contribute towards better patient outcomes.

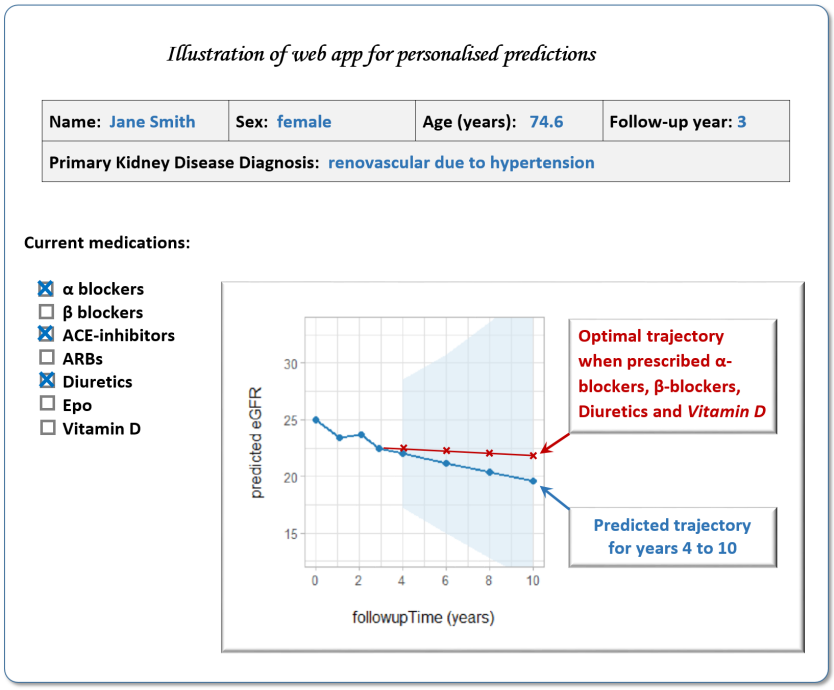

Figure 53: Illustration of web app for predicting personalised kidney disease progression

8.5.3 Treatment specific investigations

From a patient care perspective it would be useful to build further models which could give more insight into which treatments slow progression. Possible lines of inquiry are:

- A more detailed investigations with respect to our ACE-inhibitors and/or ARBs medication category; for example we could split this category in two subcategories. The SKS clinicians inform us that clinical practice assumes ACE inhibitors and ARBs will slow progression in all primary kidney disease categories. This assumption is based on two clinical studies (121) and (122) which showed improvements in renal patients with type 2 diabetes; in clinical practice it is assumed that this applies to all renal patients not just those with type 2 diabetes. However we found that ACE-inhibitors and ARBs only strongly affect progression in diabetic nephropathy. It may be that a clinical trial is needed to confirm what effect if any these drugs have on progression in other diseases.
- The SKS clinicians inform us that there is particular clinical interest is the area of anaemia management. Anaemia contributes to both poor quality of life and increases the risk of adverse outcomes, in particular cardiovascular events and death. Treatment of anaemia improves quality of life. However there is not sufficient evidence to confirm that it slows the progression of kidney disease or improves cardiovascular outcomes; for example see review (123). More detailed modelling of the SKS data could perhaps clarify this.
- It is known that EPO treatment is associated with increased risk of CV events, e.g. see review article (124). It would be of interest to determine if alternative anaemia treatments

such as intravenous iron would reduce the risk of CV events. Again more detailed modelling of the SKS data may shed some light on this.

- During early 2016 the drug Tolvaptan became available on the NHS, it is used to slow the growth of cysts in PKD patients; for example see review (125). It would therefore be of future interest, using the SKS data, to investigate the effects of this drug on the progression of PKD. We anticipate PKD patients treated with Tolvaptan will have slower kidney disease progression.

8.5.4 Further work relating to counterintuitive results

In Section 7.3 we reported several clinically counterintuitive results. After which, in Section 7.5 we took a deeper look at these results focusing on the interaction term of PTH with follow-up time. We showed there was not a clinically unexpected artefact in the SKS data relating to the slope of eGFR with respect to PTH, e.g. see Figures 48 and 49. This meant that clinical intuition matched the SKS data but not the LME model results in Section 7.3. We concluded that the counterintuitive result relating to the interaction term for PTH with follow-up time was a consequence of the way our LME model summarised the data at population level.

Future work should take a deeper look at each clinically counterintuitive result from our LME models. It should first determine if the data at individual level matches clinical intuition. If this is the case, then future work should consider alternatives to the LME models presented in this thesis. One option may be to construct a regression model which makes use of the average slope of eGFR (as it is computed in Section 4.2).

9 Conclusion

To study the key risk factors in progression of kidney disease we used data collected from over 3000 secondary care non-dialysis patients with CKD stages 3 to 5; these patients were followed-up annually until the first occurrence of one of the following end points: dialysis, kidney transplant or death. We accounted for a wide range of longitudinally recorded risk factors including comorbidities, medications, lifestyle choices, socio-demographic information and biochemical marker measurements. The role of these risk factors was considered both, between, and within, eight primary kidney disease categories including: diabetic nephropathy, glomerulonephritis, hypertensive kidney disease, pyelonephritis, renovascular disease and polycystic kidney disease. To identify key risk factors at population level we used standard longitudinal modelling techniques, in particular a linear mixed effects model with intercept and slope random effects. We robustly estimated the population level (fixed) effects in all our disease models so were able to identify key risk factors for the progression of kidney disease.

Key risk factors for lower than average levels of eGFR are biochemical markers and medications, conversely lifestyle and physical attributes are less important. More rapid progression of kidney disease is associated with biochemical markers, in contrast medications and comorbidities are not key in rapid progression. Moreover we find that PKD has the most rapid progression out of all our categories with a loss in eGFR of 3.5 mL/min/1.73m^2/year whereas the remainder have a loose around 1 mL/min/1.73m^2/year.

We suggest future work should include efforts to more cleanly record data as this could substantially improve statistical power of the statistical models. We also recommend future work should include more in depth studies of each disease category including splitting them, where appropriate, into subcategories; this would be particularly pertinent to glomerulonephritis as is contains several distinct disease types. Additionally we propose consideration should be given to using joint modelling of longitudinal and time-to-event data. Such models could be used to: study the effects of 'time-to-event' censoring; investigate treatment effects; and potentially make personalised forecasts of kidney disease progression.

We hope that this thesis research may contribute towards improvements in patient care and possibly a reduction in the burden of disease at both patient and national level.

References

1. Nahas AME, Bello AK. Chronic kidney disease: The global challenge. *The Lancet*. [Online] 2005;365(9456): 331–340. Available from: doi:10.1016/S0140-6736(05)17789-7 [Accessed: 13th June 2018]

2. Wang H, et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*. [Online] 2016;388(10053): 1459–1544. Available from: doi:10.1016/S0140-6736(16)31012-1 [Accessed: 15th June 2018]

3. Neuen BL, Chadban SJ, Demaio AR, Johnson DW, Perkovic V. Chronic kidney disease and the global NCDs agenda. *BMJ Global Health*. [Online] 2017;2(2). Available from: doi:10.1136/bmjgh-2017-000380 [Accessed: 15th June 2018]

4. Moeller S, Gioberge S, Brown G. ESRD patients in 2001: Global overview of patients, treatment modalities and development trends. *Nephrology Dialysis Transplantation*. [Online] 2002;17(12): 2071–2076. Available from: doi:10.1093/ndt/17.12.2071 [Accessed: 15th June 2018]

5. Atkins RC. The epidemiology of chronic kidney disease. *Kidney International*. [Online] 2005;67: S14–S18. Available from: doi:10.1111/j.1523-1755.2005.09403.x [Accessed: 15th June 2018]

6. World Health Organization. The top 10 global causes of death. {24 May 2018}; Available from: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Accessed: 19th June 2018]

7. World Health Organization. Global health estimates 2016: Deaths by cause, age, sex, by country and by region, 2000-2016. 2018; Available from: http://www.who.int/healthinfo/global_burden_disease/estimates/en/ [Accessed: 17th June 2018]

8. Hill NR, Fatoba ST, Oke JL, Hirst JA, O'Callaghan CA, Lasserson DS, et al. Global prevalence of chronic kidney disease – a systematic review and meta-analysis. *PLoS One*. [Online] 2016;11(7): 1–18. Available from: doi:10.1371/journal.pone.0158765 [Accessed: 11th June 2018]

9. Public Health England. Chronic kidney disease prevalence model. *PHE publications gateway number: 2014386*. [Online] 2014; Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/612303/ChronickidneydiseaseCKDprevalencemodelbriefing.pdf [Accessed: 15th June 2018]

10. NHS Digital. Quality and Outcomes Framework, 2012-13, England level: prevalence tables. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-data/ [Accessed: 20th June 2018]

11. NHS Kidney Care with Insight Health Economics. Chronic kidney disease in England: The

human and financial cost. 2012; Available from: https://www.england.nhs.uk/improvement-hub/publication/chronic-kidney-disease-in-england-the-human-and-financial-cost/ [Accessed: 13th June 2018]

12. Prakash S, O'Hare AM. Interaction of aging and CKD. *Seminars in nephrology.* [Online] 2009;29(5): 497–503. Available from: doi:10.1016/j.semnephrol.2009.06.006 [Accessed: 12th June 2018]

13. Coresh J, Selvin E, Stevens LA, Manzi J, Kusek JW, Eggers P, et al. Prevalence of chronic kidney disease in the United States. *JAMA.* [Online] 2007;298(17): 2038–2047. Available from: doi:10.1001/jama.298.17.2038 [Accessed: 12th June 2018]

14. Go AS, Chertow GM, Fan D, McCulloch CE, Hsu C-y. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *New England Journal of Medicine.* [Online] 2004;351(13): 1296–1305. Available from: doi:10.1056/NEJMoa041031 [Accessed: 13th June 2018]

15. Turin TC, James MT, Jun M, Tonelli M, Coresh J, Manns BJ, et al. Short-term change in eGFR and risk of cardiovascular events. *Journal of the American Heart Association.* [Online] 2014;3(5). Available from: doi:10.1161/JAHA.114.000997 [Accessed: 22nd June 2018]

16. Rifkin DE, Shlipak MG, Katz R, Fried LF, Siscovick D, Chonchol M, et al. Rapid kidney function decline and mortality risk in older adults. *Archives of Internal Medicine.* [Online] 2008;168(20): 2212–2218. Available from: doi:10.1001/archinte.168.20.2212 [Accessed: 22nd June 2018]

17. Liu M, Li XC, Lu L, Cao Y, Sun RR, Chen S, et al. Cardiovascular disease and its relationship with chronic kidney disease. *European Review for Medical and Pharmacological Sciences.* [Online] 2014;18(19): 2918–2926. Available from: https://www.europeanreview.org/article/7900 [Accessed: 13th June 2018]

18. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD. United States Renal Data System annual data report: Epidemiology of kidney disease in the United States. Volume 1, Chronic Kidney Disease in the United States. 2017; Available from: https://www.usrds.org/2017/view/ [Accessed: 13th June 2018]

19. Thomas R, Kanso A, Sedor JR. Chronic kidney disease and its complications. *Primary Care.* [Online] 2008;35(2): 329–344. Available from: doi:10.1016/j.pop.2008.01.008 [Accessed: 13th June 2018]

20. Hernandez GT, Nasri H. World Kidney Day 2014: Increasing awareness of chronic kidney disease and aging. *Journal of renal injury prevention.* [Online] 2014;3(1): 3–4. Available from: doi:10.12861/jrip.2014.02 [Accessed: 15th June 2018]

21. Levey AS, Astor BC, Stevens LA, Coresh J. Chronic kidney disease, diabetes, and hyper-

tension: What's in a name? *Kidney International*. [Online] 2010;78(1): 19–22. Available from: doi:https://doi.org/10.1038/ki.2010.115 [Accessed: 15th June 2018]

22. Gullion CM, Keith DS, Nichols GA, Smith DH. Impact of comorbidities on mortality in managed care patients with CKD. *American Journal of Kidney Diseases*. [Online] 2006;48(2): 212–220. Available from: doi:10.1053/j.ajkd.2006.04.083 [Accessed: 13th June 2018]

23. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD. United States Renal Data System annual data report: Epidemiology of kidney disease in the United States. Volume 2, End-stage Renal Disease (ESRD) in the United States. 2017; Available from: https://www.usrds.org/2017/view/ [Accessed: 13th June 2018]

24. James MT, Hemmelgarn BR, Tonelli M. Early recognition and prevention of chronic kidney disease. *The Lancet*. [Online] 2010;375(9722): 1296–1309. Available from: doi:10.1016/S0140-6736(09)62004-3 [Accessed: 15th June 2018]

25. Snively CS, Gutierrez C. Chronic kidney disease: Prevention and treatment of common complications. *American Family Physician*. [Online] 2004;70(10): 1921–1928. Available from: https://www.aafp.org/afp/2004/1115/p1921.html [Accessed: 14th June 2018]

26. Saweirs WWM, Goddard J. What are the best treatments for early chronic kidney disease? A background paper prepared for the UK Consensus Conference on Early Chronic Kidney Disease. *Nephrology Dialysis Transplantation*. [Online] 2007;22(suppl_9): ix31–ix38. Available from: doi:10.1093/ndt/gfm447 [Accessed: 15th June 2018]

27. National Kidney Foundation. GFR (Glomerular Filtration Rate): A key to understanding how well your kidneys are working. *Nephrology Dialysis Transplantation*. [Online] 2013; Available from: https://www.kidney.org/sites/default/files/docs/11-10-1813_abe_patbro_gfr_b.pdf [Accessed: 18th June 2018]

28. Delanaye P, Schaeffner E, Ebert N, Cavalier E, Mariat C, Krzesinski J-M, et al. Normal reference values for glomerular filtration rate: What do we really know? *Nephrology Dialysis Transplantation*. [Online] 2012;27(7): 2664–2672. Available from: doi:10.1093/ndt/gfs265 [Accessed: 18th June 2018]

29. Sheen Y-J, Sheu WH. Risks of rapid decline renal function in patients with type 2 diabetes. *World Journal of Diabetes*. [Online] 2014;5(6): 835–846. Available from: doi:10.4239/wjd.v5.i6.835 [Accessed: 18th June 2018]

30. National Institute for Health and Care Excellence (NICE). Chronic kidney disease in adults: Assessment and management. *NICE clinical guideline CG182*. [Online] 2014; Available from: https://www.nice.org.uk/guidance/cg182 [Accessed: 15th June 2018]

31. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney

disease. *Kidney International*. [Online] Agency for Healthcare Research; Quality (AHRQ); 2013;3(1): 1–150. Available from: https://guideline.gov/summaries/summary/46510/kdigo-2012-clinical-practice-guideline-for-the-evaluation-and-management-of-chronic-kidney-disease/ [Accessed: 8th June 2018]

32. Ali O, Mohiuddin A, Mathur R, Dreyer G, Hull S, Yaqoob MM. A cohort study on the rate of progression of diabetic chronic kidney disease in different ethnic groups. *BMJ Open*. [Online] 2013;3(2). Available from: doi:10.1136/bmjopen-2012-001855 [Accessed: 21st June 2018]

33. Higashihara E, Horie S, Muto S, Mochizuki T, Nishio S, Nutahara K. Renal disease progression in autosomal dominant polycystic kidney disease. *Clinical and Experimental Nephrology*. [Online] 2012;16(4): 622–628. Available from: doi:10.1007/s10157-012-0611-9 [Accessed: 21st June 2018]

34. Hoefield RA, Kalra PA, Baker P, Lane B, New JP, O'Donoghue DJ, et al. Factors associated with kidney disease progression and mortality in a referred CKD population. *American Journal of Kidney Diseases*. [Online] 2010;56(6): 1072–1081. Available from: doi:10.1053/j.ajkd.2010.06.010 [Accessed: 16th June 2018]

35. Eddington H, Hoefield R, Sinha S, Chrysochou C, Lane B, Foley RN, et al. Serum phosphate and mortality in patients with chronic kidney disease. *Clinical Journal of the American Society of Nephrology*. [Online] 2010;5(12): 2251–2257. Available from: doi:10.2215/CJN.00810110 [Accessed: 16th June 2018]

36. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. [Online] Springer-Verlag, New-York; 2000. Available from: doi:10.1007/978-1-4419-0300-6 [Accessed: 4th July 2018]

37. Hedeker D, Gibbons RD. *Longitudinal data analysis*. [Online] John Wiley & Sons, Hoboken; 2006. Available from: doi:10.1002/0470036486 [Accessed: 4th July 2018]

38. Fitzmaurice G, Laird NM, Ware JH. *Applied longitudinal analysis. 2nd edition*. John Wiley & Sons, Hoboken; 2011.

39. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. [Online] 1982;38(4): 963–974. Available from: doi:10.2307/2529876 [Accessed: 4th July 2018]

40. Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics*. [Online] 1988;44(4): 959–971. Available from: https://doi.org/10.2307/2531727 [Accessed: 18th September 2019]

41. Boucquemont J, Heinze G, Jager KJ, Oberbauer R, Leffondré K. Regression methods for investigating risk factors of chronic kidney disease outcomes: The state of the art. *BMC Nephrology*. [Online] 2014;15(45). Available from: https://doi.org/10.1186/1471-2369-15-45

[Accessed: 18th September 2019]

42. Janmaat CJ, Diepen M van, Tsonaka R, Jager KJ, Zoccali C, Dekker FW. Pitfalls of linear regression for estimating slopes over time and how to avoid them by using linear mixed-effects models. *Nephrology Dialysis Transplantation*. [Online] 2018;34(4): 561–566. Available from: doi:10.1093/ndt/gfy128 [Accessed: 18th September 2019]

43. Shou H, Hsu JY, Xie D, Yang W, Roy J, Anderson AH, et al. Analytic considerations for repeated measures of eGFR in cohort studies of CKD. *Clinical journal of the American Society of Nephrology : CJASN*. [Online] 2017;12(8). Available from: https://doi.org/10.2215/CJN.11311116 [Accessed: 18th September 2019]

44. Shou H, Hsu JY, Xie D, Yang W, Roy J, Anderson AH, et al. Supplemental Material: Analytic considerations for repeated measures of eGFR in cohort studies of CKD. *Clinical journal of the American Society of Nephrology : CJASN*. [Online] 2017;12(8). Available from: https://cjasn.asnjournals.org/content/suppl/2017/07/27/CJN.11311116.DCSupplemental [Accessed: 18th September 2019]

45. Leffondre K, Boucquemont J, Tripepi G, Stel VS, Heinze G, Dunkler D. Analysis of risk factors associated with renal function trajectory over time: A comparison of different statistical approaches. *Nephrology Dialysis Transplantation*. [Online] 2014;30(8): 1237–1243. Available from: doi:https://doi.org/10.1093/ndt/gfu320 [Accessed: 18th September 2019]

46. R Core Team. *R: A language and environment for statistical computing*. [Online] Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: http://www.R-project.org/ [Accessed: 5th June 2018]

47. National Institute for Health and Care Excellence (NICE). Obesity: Identification, assessment and management. *NICE clinical guideline CG189*. [Online] 2014; Available from: https://www.nice.org.uk/guidance/cg189 [Accessed: 25th June 2018]

48. Asmar R, Vol S, Brisac A-M, Tichet J, Topouchian J. Reference values for clinic pulse pressure in a nonselected population. *American Journal of Hypertension*. [Online] 2001;14(5): 415–418. Available from: doi:10.1016/S0895-7061(01)01284-5 [Accessed: 27th June 2018]

49. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*. [Online] 2009;150(9): 604–612. Available from: doi:10.7326/0003-4819-150-9-200905050-00006 [Accessed: 7th June 2018]

50. Levey AS, Stevens LA. Estimating GFR using the CKD Epidemiology Collaboration (CKD-EPI) creatinine equation: More accurate GFR estimates, lower CKD prevalence estimates, and better risk predictions. *American Journal of Kidney Diseases*. [Online] 2010;55(4): 622–627.

Available from: doi:10.1053/j.ajkd.2010.02.337 [Accessed: 22nd June 2018]

51. Michels WM, Grootendorst DC, Verduijn M, Elliott EG, Dekker FW, Krediet RT. Performance of the Cockcroft-Gault, MDRD, and new CKD-EPI formulas in relation to GFR, age, and body size. *Clinical Journal of the American Society of Nephrology.* [Online] 2010;5(6): 1003–1009. Available from: doi:10.2215/CJN.06870909 [Accessed: 22nd June 2018]

52. Seegmiller JC, Eckfeldt JH, Lieske JC. Challenges in measuring glomerular filtration rate: A clinical laboratory perspective. *Advances in Chronic Kidney Disease.* [Online] 2018;25(1): 84–92. Available from: doi:10.1053/j.ackd.2017.10.006 [Accessed: 22nd June 2018]

53. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika.* [Online] 1965;52(3-4): 591–611. Available from: doi:10.1093/biomet/52.3-4.591 [Accessed: 30th July 2018]

54. Kilbride HS, Stevens PE, Eaglestone G, Knight S, Carter JL, Delaney MP, et al. Accuracy of the MDRD (Modification of Diet in Renal Disease) Study and CKD-EPI (CKD Epidemiology Collaboration) Equations for Estimation of GFR in the Elderly. *American Journal of Kidney Diseases.* [Online] 2013;61(1): 57–66. Available from: doi:10.1053/j.ajkd.2012.06.016 [Accessed: 2nd May 2019]

55. Rubin DB. *Multiple imputation for nonresponse in surveys.* [Online] John Wiley & Sons, New York; 1987. Available from: doi:10.1002/9780470316696 [Accessed: 19th July 2018]

56. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association.* [Online] 1996;91(434): 473–489. Available from: doi:10.1080/01621459.1996.10476908 [Accessed: 19th July 2018]

57. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* [Online] 1977;39(1): 1–38. Available from: http://www.jstor.org/stable/2984875 [Accessed: 19th July 2018]

58. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software.* [Online] 2011;45(3): 1–67. Available from: doi:10.18637/jss.v045.i03 [Accessed: 26th April 2019]

59. Honaker J, King G, Blackwell M. Amelia II: A program for missing data. *Journal of Statistical Software.* [Online] 2011;45(7): 1–47. Available from: doi:10.18637/jss.v045.i07 [Accessed: 19th July 2018]

60. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork J. Comparison of different methods for univariate time series imputation in R. *ArXiv e-prints.* [Online] 2015; Available from:

https://arxiv.org/abs/1510.03924 [Accessed: 19th July 2018]

61. Moritz S, Bartz-Beielstein T. imputeTS: Time series missing value imputation in R. *The R Journal*. [Online] 2017;9(1): 207–218. Available from: https://journal.r-project.org/archive/2017/RJ-2017-009/ [Accessed: 16th July 2018]

62. Welch G, Bishop G. An introduction to the Kalman filter. *ACM SIGGRAPH 2001*. [Online] 2001; Available from: http://www.cs.unc.edu/~tracker/media/pdf/SIGGRAPH2001_CoursePack_08.pdf [Accessed: 20th July 2018]

63. Maybeck PS. *Stochastic models, estimation, and control. Volume 1*. Academic Press, New York, San Francisco, London; 1979.

64. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. [Online] 1986;73(1): 13–22. Available from: doi:10.1093/biomet/73.1.13 [Accessed: 30th January 2019]

65. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of longitudinal data. 2nd edition*. [Online] Oxford University Press, Oxford; 2013. Available from: http://www.oup.co.uk/isbn/0-19-852484-6 [Accessed: 4th July 2018]

66. Pinheiro JC, Bates DM. *Mixed-effects Models in S and S-PLUS*. Springer, New York; 2000.

67. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. *nlme: Linear and nonlinear mixed effects models*. [Online] 2018. Available from: https://CRAN.R-project.org/package=nlme [Accessed: 27th August 2018]

68. Chatterjee S, Hadi AS. *Regression analysis by example. 4th edition*. [Online] John Wiley & Sons, Hoboken; 2006. Available from: doi:10.1002/0470055464

69. Naimi B, Hamm NAS, Groen TA, Skidmore AK, Toxopeus AG. Where is positional uncertainty a problem for species distribution modelling? *Ecography*. [Online] 2014;37(2): 191–203. Available from: doi:10.1111/j.1600-0587.2013.00205.x [Accessed: 7th August 2018]

70. O'brien RM. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*. [Online] 2007;41(5): 673–690. Available from: doi:10.1007/s11135-006-9018-6 [Accessed: 11th August 2018]

71. Akaike H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory. Budapest*. 1973; 267–281.

72. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. [Online] 1974;19(6): 716–723. Available from: doi:10.1109/TAC.1974.1100705 [Accessed:

17th August 2018]

73. Akaike H. A new look at the statistical model identification. *Current Contents Engineering, Technology, and Applied Sciences.* [Online] 1981;12(51): 42. Available from: http://www.garfield.library.upenn.edu/classics1981/A1981MS54100001.pdf [Accessed: 17th August 2018]

74. Schwarz GE. Estimating the dimension of a model. *Annals of Statistics.* [Online] 1978;6(2): 461–464. Available from: doi:10.1214/aos/1176344136 [Accessed: 17th August 2018]

75. Vandekerckhove J, Matzke D, Wagenmakers E-J. *Model comparison and the principle of parsimony. The Oxford handbook of computational and mathematical psychology.* [Online] Oxford University Press, New York; 2015. Available from: http://www.ejwagenmakers.com/inpress/VandekerckhoveEtAlinpress.pdf. Available from: doi:10.1093/oxfordhb/9780199957996.013.14 [Accessed: 17th August 2018]

76. Yang Y. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika.* [Online] 2005;92(4): 937–950. Available from: doi:10.1093/biomet/92.4.937 [Accessed: 18th August 2018]

77. Vrieze SI. Model selection and psychological theory: A discussion of the differences between the akaike information criterion (AIC) and the bayesian information criterion (BIC). *Psychological Methods.* [Online] 2012;17(2): 228–243. Available from: doi:10.1037/a0027127 [Accessed: 17th August 2018]

78. Venables WN, Ripley BD. *Modern applied statistics with S. 4th edition.* [Online] Springer, New York; 2002. Available from: http://www.stats.ox.ac.uk/pub/MASS4 [Accessed: 16th August 2018]

79. Efron B. Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics.* [Online] 1979;7(1): 1–26. Available from: doi:10.1214/aos/1176344552 [Accessed: 23rd August 2018]

80. Efron B, Tibshirani RJ. *An introduction to the bootstrap.* Chapman & Hall, New York; 1993.

81. Chernick MR, LaBudde RA. *An introduction to bootstrap methods with applications to R.* John Wiley & Sons, Hoboken; 2011.

82. Bickel PJ, Götze F, van Zwet WR. Resampling fewer than n observations: Gains, losses and remedies for losses. *Statistica Sinica.* [Online] 1997;7: 1–31. Available from: http://www3.stat.sinica.edu.tw/statistica/oldpdf/a7n11.pdf [Accessed: 24th August 2018]

83. Canty A, Ripley BD. *boot: Bootstrap R (S-Plus) Functions.* [Online] 2019. Available from: https://cran.r-project.org/web/packages/boot/ [Accessed: 13th September 2019]

84. Davison AC, Hinkley DV. *Bootstrap methods and their applications.* [Online] Cambridge

University Press, Cambridge; 1997. Available from: http://statwww.epfl.ch/davison/BMA/ [Accessed: 13th September 2019]

85. Hoefield RA, Kalra PA, Lane B, O'Donoghue DJ, Foley RN, Middleton RJ. Associations of baseline characteristics with evolution of eGFR in a referred chronic kidney disease cohort. *QJM: An International Journal of Medicine.* [Online] 2013;106(10): 915–924. Available from: doi:10.1093/qjmed/hct115 [Accessed: 28th February 2019]

86. Feldman HI, Appel LJ, Chertow GM, Cifelli D, Cizman B, Daugirdas J, et al. The Chronic Renal Insufficiency Cohort (CRIC) Study: Design and Methods. *Journal of the American Society of Nephrology.* [Online] American Society of Nephrology; 2003;14(suppl 2): S148–S153. Available from: doi:10.1097/01.ASN.0000070149.78399.CE [Accessed: 2nd May 2019]

87. The Chronic Renal Insufficiency Cohort Study. 2019; Available from: http://www.cristudy.org/ [Accessed: 2nd May 2019]

88. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research.* [Online] 2011;20(1): 40–49. Available from: doi:10.1002/mpr.329 [Accessed: 26th April 2019]

89. Flom PL, Cassell DL. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. *NESUG.* [Online] 2007; 1–7. Available from: https://www.lexjansen.com/pnwsug/2008/DavidCassell-StoppingStepwise.pdf [Accessed: 24th April 2019]

90. Chatfield C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A.* [Online] 1995;158(3): 419–466. Available from: doi:10.2307/2983440 [Accessed: 25th April 2019]

91. Harrell FE. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. 2nd edition.* Springer-Verlag, New York; 2015.

92. Emmert-Streib F, Dehmer M. Evaluation of regression models: Model assessment, model selection and generalization error. *Machine learning and knowledge extraction.* [Online] 2019;1: 521–550. Available from: doi:10.3390/make1010032

93. Wu L. *Mixed effects models for complex data (monographs on statistics and applied probability book 113).* Chapman & Hall/CRC, Boca Raton; 2009.

94. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: A mediation analysis. *Social Science & Medicine.* [Online] 2017;195: 42–49. Available from: doi:10.1016/j.socscimed.2017.11.008 [Accessed: 20th September 2019]

95. Kiecolt-Glaser JK, McGuire L, Robles TF, Glaser R. Emotions, morbidity, and mortality: New perspectives from psychoneuroimmunology. *Annual Review of Psychology.* [Online] 2002;53(1):

83–107. Available from: doi:10.1146/annurev.psych.53.100901.135217 [Accessed: 20th September 2019]

96. Scott KM, Bruffaerts R, Tsang A, Ormel J, Alonso J, Angermeyer MC, et al. Depression anxiety relationships with chronic physical conditions: Results from the World Mental Health surveys. *Journal of Affective Disorders*. [Online] 2007;103(1): 113–120. Available from: doi:10.1016/j.jad.2007.01.015 [Accessed: 20th September 2019]

97. Bonneta F, Irvinga K, Terrab J-L, Nonyc P, Berthezènea F, Moulina P. Anxiety and depression are associated with unhealthy lifestyle in patients at risk of cardiovascular disease. *Atherosclerosis*. [Online] 2005;178(2): 339–344. Available from: doi:10.1016/j.atherosclerosis.2004.08.035 [Accessed: 20th September 2019]

98. Hosmer DW, Lemeshow S, May S. *Applied survival analysis: Regression modeling of time-to-event data. 2nd edition*. [Online] John Wiley & Sons, Hoboken; 2008. Available from: doi:10.1002/9780470258019 [Accessed: 9th May 2019]

99. Schober P, Vetter T. Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia and Analgesia*. [Online] 2018;127(3): 792–798. Available from: doi:10.1213/ANE.0000000000003653 [Accessed: 10th May 2019]

100. Junyong I, Kyu LD. Survival analysis: Part I - analysis of time-to-event. *Korean Journal of Anesthesiology*. [Online] 2018;71(3): 182–191. Available from: doi:10.4097/kja.d.18.00067 [Accessed: 10th May 2019]

101. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society - Statistical Methodology*. [Online] 1972;34(2): 187–220. Available from: http://www.jstor.org/stable/2985181 [Accessed: 9th May 2019]

102. Cox DR. Partial likelihood. *Biometrika*. [Online] 1975;62(2): 269–276. Available from: doi:10.1093/biomet/62.2.269 [Accessed: 9th May 2019]

103. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: Further concepts and methods in survival analysis. *British journal of cancer*. [Online] 2003;89(5): 781–786. Available from: doi:10.1038/sj.bjc.6601117 [Accessed: 10th May 2019]

104. Altman DG, De Stavola BL. Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Statistics in Medicine*. [Online] 1994;13(4): 301–341. Available from: doi:10.1002/sim.4780130402 [Accessed: 10th May 2019]

105. Asar Ö, Ritchie J, Kalra PA, Diggle PJ. Joint modelling of repeated measurement and time-to-event data: An introductory tutorial. *International Journal of Epidemiology*. [Online]

2015;44(1): 334–344. Available from: doi:10.1093/ije/dyu262 [Accessed: 6th May 2019]

106. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *American Journal of Epidemiology*. [Online] 2009;170(2): 244–256. Available from: doi:10.1093/aje/kwp107 [Accessed: 9th May 2019]

107. Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation*. [Online] 2013;28(11): 2670–2677. Available from: doi:10.1093/ndt/gft355 [Accessed: 9th May 2019]

108. Williamson PR, Kolamunnage-Dona R, Philipson P, Marson AG. Joint modelling of longitudinal and competing risks data. *Statistics in Medicine*. [Online] 2008;27(30): 6426–6438. Available from: doi:10.1002/sim.3451 [Accessed: 10th May 2019]

109. Berlin KS, Williams NA, Parra GR. An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology*. [Online] 2013;39(2): 174–187. Available from: doi:10.1093/jpepsy/jst084 [Accessed: 8th January 2019]

110. Berlin KS, Parra GR, Williams NA. An introduction to latent variable mixture modeling (part 2): Longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology*. [Online] 2013;39(2): 188–203. Available from: doi:10.1093/jpepsy/jst085 [Accessed: 8th January 2019]

111. Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*. [Online] 2017;78(2): 1–56. Available from: doi:10.18637/jss.v078.i02 [Accessed: 8th January 2019]

112. Proust-Lima C, Philipps V, Diakite A, Liquet B. *Lcmm: Extended mixed models using latent classes and latent processes*. [Online] 2018. Available from: https://cran.r-project.org/package=lcmm [Accessed: 8th January 2019]

113. NHS England. Improving outcomes through personalised medicine. 2016; Available from: https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf [Accessed: 2nd May 2019]

114. Jakka S, Rossbach M. An economic perspective on personalized medicine. *The HUGO Journal*. [Online] 2013;7(1): 1. Available from: doi:10.1186/1877-6566-7-1 [Accessed: 6th May 2019]

115. Personalized Medicine Coalition, Washington. The case for personalized medicine. 4th edition. 2014; 1–68. Available from: http://www.personalizedmedicinecoalition.org/Userfiles/

PMC-Corporate/file/pmc_case_for_personalized_medicine.pdf [Accessed: 6th May 2019]

116. Mathur S, Sutton J. Personalized medicine could transform healthcare (Review). *Biomedical Reports*. [Online] 2017;7(1): 3–5. Available from: doi:10.3892/br.2017.922 [Accessed: 6th May 2019]

117. Diggle PJ, Sousa I, Asar Ö. Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*. [Online] 2015;16(3): 522–536. Available from: doi:10.1093/biostatistics/kxu053 [Accessed: 31st August 2018]

118. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology*. [Online] 2016;16(117): 1–15. Available from: doi:10.1186/s12874-016-0212-5 [Accessed: 1st May 2019]

119. Brankovic M, Kardys I, Hoorn EJ, Baart S, Boersma E, Rizopoulos D. Personalized dynamic risk assessment in nephrology is a next step in prognostic research. *Kidney International*. [Online] 2018;94(1): 214–217. Available from: doi:10.1016/j.kint.2018.04.007 [Accessed: 6th May 2019]

120. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: Data science enabling personalized medicine. *BMC Medicine*. [Online] 2018;16(1): 150. Available from: doi:10.1186/s12916-018-1122-7 [Accessed: 6th May 2019]

121. Lewis EJ, Hunsicker LG, Clarke WR, Berl T, Pohl MA, Lewis JB, et al. Renoprotective effect of the angiotensin-receptor antagonist Irbesartan in patients with nephropathy due to type 2 diabetes. *New England Journal of Medicine*. [Online] 2001;345(12): 851–860. Available from: doi:10.1056/NEJMoa011303 [Accessed: 15th October 2019]

122. Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, Parving H-H, et al. Effects of Losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *New England Journal of Medicine*. [Online] 2001;345(12): 861–869. Available from: doi:10.1056/NEJMoa011161 [Accessed: 15th October 2019]

123. Mehdi U, Toto RD. Anemia, diabetes, and chronic kidney disease. *Diabetes Care*. [Online] 2009;32(7): 1320–1326. Available from: doi:10.2337/dc08-0779 [Accessed: 15th October 2019]

124. Provatopoulou S, Ziroyiannis P. Clinical use of erythropoietin in chronic kidney disease: Outcomes and future prospects. *Hippokratia*. [Online] 2011;15(2): 109–115. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3208971/ [Accessed: 15th October 2019]

125. Mustafa RA, Yu ASL. Burden of proof for Tolvaptan in ADPKD. *Clinical Journal of the American Society of Nephrology*. [Online] 2018;13(7): 1107–1109. Available from:

doi:10.2215/CJN.00190118 [Accessed: 15th October 2019]

126. Department of Health and Social Care. UK Chief Medical Officers' low risk drinking guidelines. 2016; Available from: http://www.gov.uk/government/publications/alcohol-consumption-advice-on-low-risk-drinking [Accessed: 12th July 2018]

127. Office for National Statistics. Standard Occupational Classification 2010 Volume 3: The National Statistics Socio-economic classification (NS-SEC rebased on the SOC2010). ISBN:978-0-230-27224-8. Palgrave Macmillan, Basingstoke, UK; 2010; Available from: https://www.ons.gov.uk/methodology/classificationsandstandards/ [Accessed: 12th July 2018]

Appendix

A.1 Data cleaning and preparation

We prepared and cleaned the SKS data as follows:

- Fixed common mis-spellings and removed erroneous characters in fields.
- All measurements were converted so they had consistent units
- Matched follow-up appointments with biochemical records by finding closest biochemical date to follow-up date. Generally the mismatch between dates was less than a few days. When date difference was larger than six weeks we assumed there was no match.
- At each consultation two measurements of systolic and diastolic were recorded. We subtracted mean systolic from diastolic blood pressure to compute mean pulse pressure.
- After converting units to kilograms and metres BMI was computed as weight/(height)^2.
- eGFR was calculated using Equation 1 after ensuring consistent units.
- The urine protein rate (mg/24hr) and PCR (g/mol) measurements were combined into a single quantity 'proteinuria' (mg/24hr); note for example PCR 50mg/mmol=50g/mol is equivalent to urinary protein rate 0.5g/24hr=500mg/24hr.
- HbA1c measured in %Hb was converted to mmol/mol, where HbA1c(mmol/mol) = 10.929 $\times$ (HbA1c(%Hb) - 2.15).
- Units of alcohol per week were summarised into three categories: less than 1, 1 to 14 and over 14. We defined these categories to be in-line with UK Chief Medical Officers' guidelines issued during 2016 (126).
- Smoking status was defined as: ex-smoker, active smoker, non-smoker.
- Ethnicity was defined as follows: white, Asian, black, Chinese, other. Of these categories the last four contain small numbers of patients so they were grouped into a single category labelled 'non-white'.
- The primary occupation of patients, in some cases prior to retirement, was recorded by SKS using the 8 classes defined by the Office of National Statistics (ONS) under their socio-economic classification (127). To increase our statistical power we reduced the number of occupation classes using the ONS guidelines in Section 7 *Classes and collapses* of (127). The categories we used are:
  - Higher managerial, administrative and professional occupations (abbreviation *ManagerialProfessional*)
  - Intermediate occupations (abbreviation *Intermediate*)
  - Routine and manual occupations (abbreviation *RoutineManual*)
  - Never worked and long-term unemployed (abbreviation *NeverWorkedUnemployed*)
- AKI episodes were identified as at least 3 days of consecutive creatinine measurements; this implied the patient had been admitted to hospital.

A.1.1 End of study markers

If a patient withdraw from the study, and the reason was recorded, these reasons were defined as either lost-to-follow-up, death or RRT. In this thesis we added two additional categories; *presumed lost-to-follow-up* and *ongoing* to capture patients with no recorded reason for leaving. For these patients, if their last follow-up date was more than than 2 years 6 months before 27 February 2017 (most recent date in data), then we recorded them as *presumed lost-to-follow-up*, otherwise we assumed they were still in the study so defined them as *ongoing*. This cut-off was set to over 2.5 years because some of the cohort had follow-ups which were two years apart and moreover follow-ups were not separated by precisely one or two calendar years. The date of departure from the study is usually only known/recorded for patients who die or undergo RRT.

Note that if follow-ups are two years apart then their annual follow-up index skips a year. For example, if a patient is seen at follow-up '1' (first follow-up) and then is next seen two years later their follow-up index jumps to '3', i.e. '2' is skipped.

A.1.2 Primary kidney disease categories

Kidney diseases were grouped into 9 primary disease categories:

- *diabetic nephropathy*: single disease - no subgroups.
- *glomerulonephritis*: crescentic and focal segmental glomerulonephritis, Goodpasture's syndrome, Henoch-Schonlein purpura, IgA nephropathy, Lupus erythematosus, membranoproliferative, membranous nephropathy, Wegener's Granulomatosis, and renal vascular disease due to polyarteritis.
- *hypertensive kidney disease*: renal vascular disease due to hypertension or due to malignant hypertension, and ischemic renal disease / cholesterol embolism (standard diagnostic (EDTA) codes 71,72 and 75).
- *obstruction*: obstructive uropathy - no subgroups.
- *other*: kidney disease which does not come under any of the other 8 categories.
- *polycystic kidney disease*: single disease - no subgroups.
- *pyelonephritis*, due to obstructive uropathy, urolithiasis, vesico-ureteric reflux without obstruction, associated with neurogenic bladder and other cause.
- *renovascular*, due to polyarteritis and other reason.
- *unknown*: type of renal disease not diagnosed.

A very small minority of patients had more than one primary kidney disease recorded. Typically early during their time in the study they were assigned a diagnosis of 'unknown' then later given a diagnosis from one of the other categories. In such cases we assigned their primary kidney disease as the diagnosed disease hence in this thesis each patient only has one primary kidney disease type.

A.1.3 Comorbidity categories

Comorbidities were collated into 5 groups as follows:

- *cardivascular*: amputation, angina, cardiac arrest, cerebrovascular disease, coronary intervention, heart failure, myocardial infarction, peripheral vascular disease
- *gastrointestinal*: peptic ulcer disease, liver disease
- *cancer*: any type of cancer, not necessarily kidney cancer, which may either be 'current' or 'previous'
- *diabetes*: type 1 and 2
- *other*: parathyroidectomy, dementia, chronic obstructive pulmonary disease, congenital abnormalities

A.1.4 Medication categories

Drugs were categorised as follows:

- *ACE inhibitor*: captopril, cilazapril, enalapril, fosinopril, imidapril, lisinopril, perindopril, quinapril, ramipril, trandolapril
- *ARB*: candesartan, eprosartan, irbesartan, losartan, olmesartan, telmisartan, valsartan
- *alpha-blocker*: alfuzosin, doxazosin, indoramin, mirtazapine, prazosin, tamsulosin, terazosin, trazodone, yohimbine
- *beta-blocker*: acebutalol, atenolol, betaxolol, bisoprolol, celiprolol, metoprolol, nebivolol, propranolol, sotalol, timolol
- *combined alpha- and beta-blocker*: carvedilol, labetalol
- *calcium channel blocker*: amlodipine, diltiazem, coracten, felodipine, lacidipine, lercanidipine, nicardipine, nifedipine, securon, verapamil
- *diuretic*: amiloride, bendroflumethiazide, bendrofluazide, bumetanide, chlorthalidone, chlorothiazide, co-amilozide, co-amilofruse, eplerenone, frusemide, hydrochlorothiazide, indapamide, metolazone, spironolactone, thiazide, torasemide, xipamide
- *EPO*: treatment using EPO was originally recorded in one of three groups (epoetin alpha, beta or darbepoetin), here we group them in a single category.
- *iron* (taken orally): ferrous sulphate, ferrous fumarate, iron, ferrous gluconate, iron sulphate, ferrograd, fersamal
- *vitamin D*: alfacalcidol, cholecalciferol, vitamin D

The SKS data records each patient as either taking or not taking parenteral iron. The drugs administered are not recorded.

A.2 Trellis plots of eGFR against follow-up for individual patients

Each figure contains Trellis plots based on the first 24 patients in the SKS data with a given primary disease. The points on each plot are log(eGFR) at each follow-up year for which data were recorded; a maximum of 10 years are plotted. Superimposed on each plot is a straight line derived from a linear model for the given patient where the outcome variable is log(eGFR) and the covariate is follow-up year; this model uses eGFR values from all follow-up years hence is not truncated at 10 years.

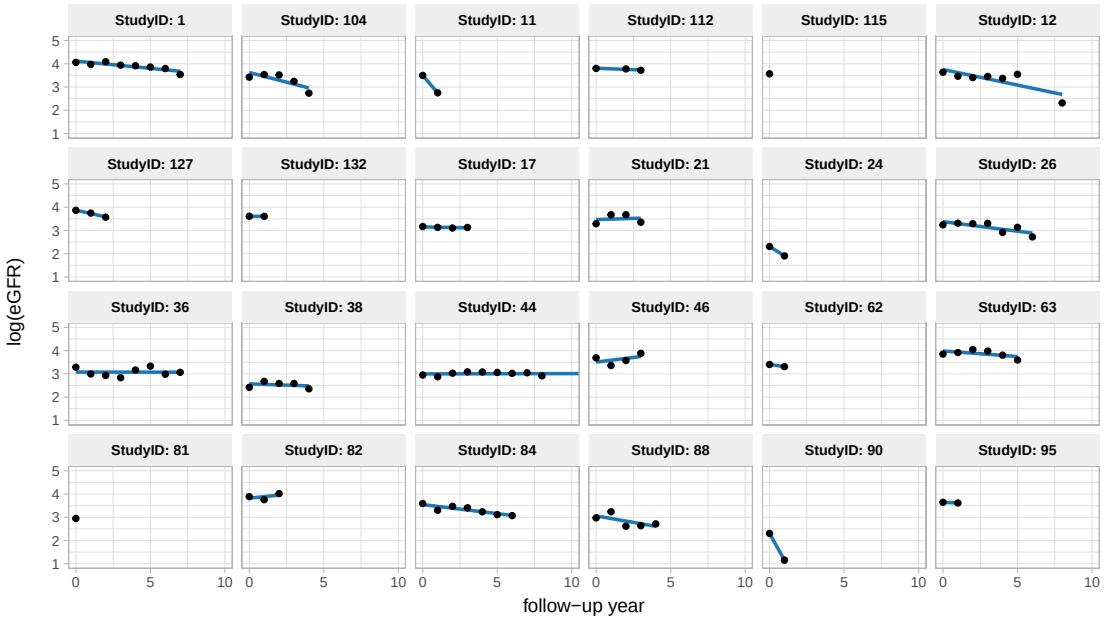

Figure 54: Progression of disease for 24 patients with diabetic nephropathy

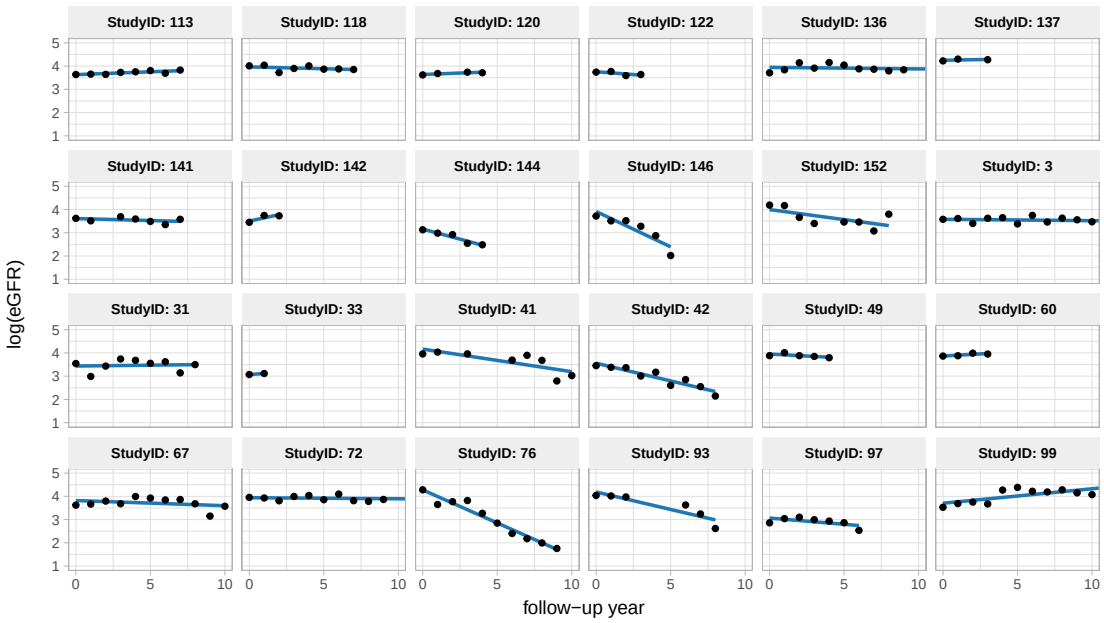

Figure 55: Progression of disease for 24 patients with glomerulonephritis

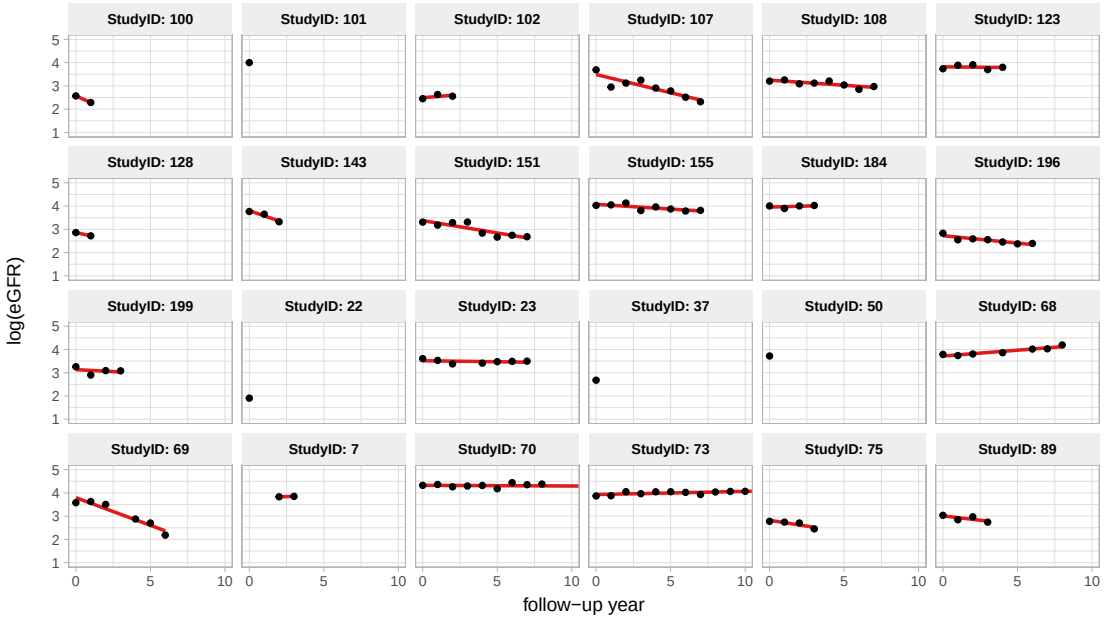

Figure 56: Progression of disease for 24 patients with HKD

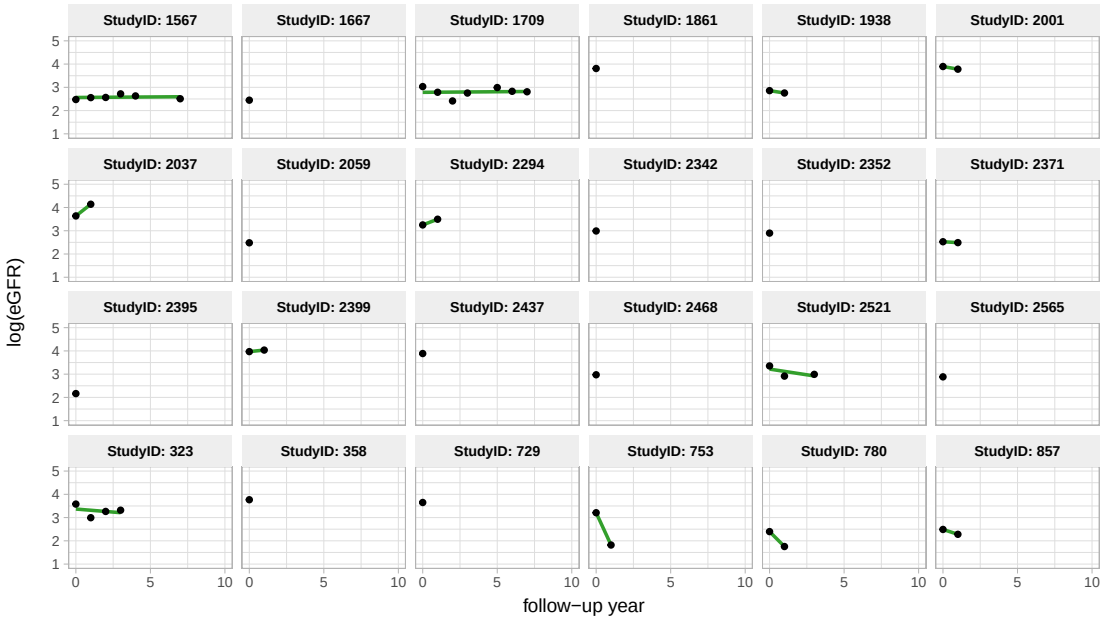

Figure 57: Progression of disease for 24 patients with obstruction

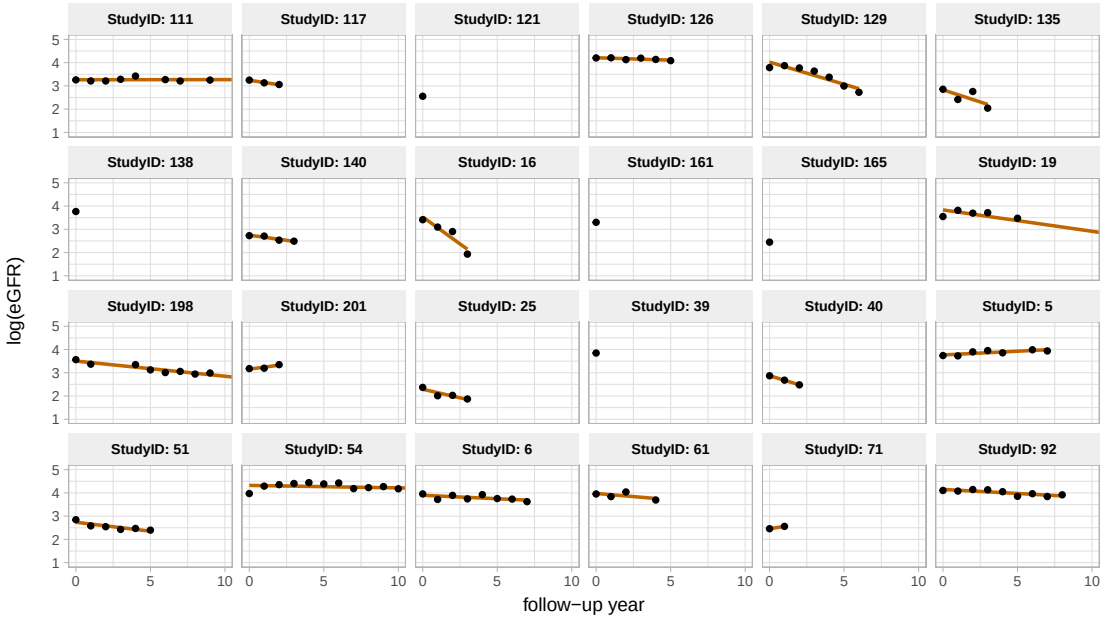

Figure 58: Progression of disease for 24 patients with disease other

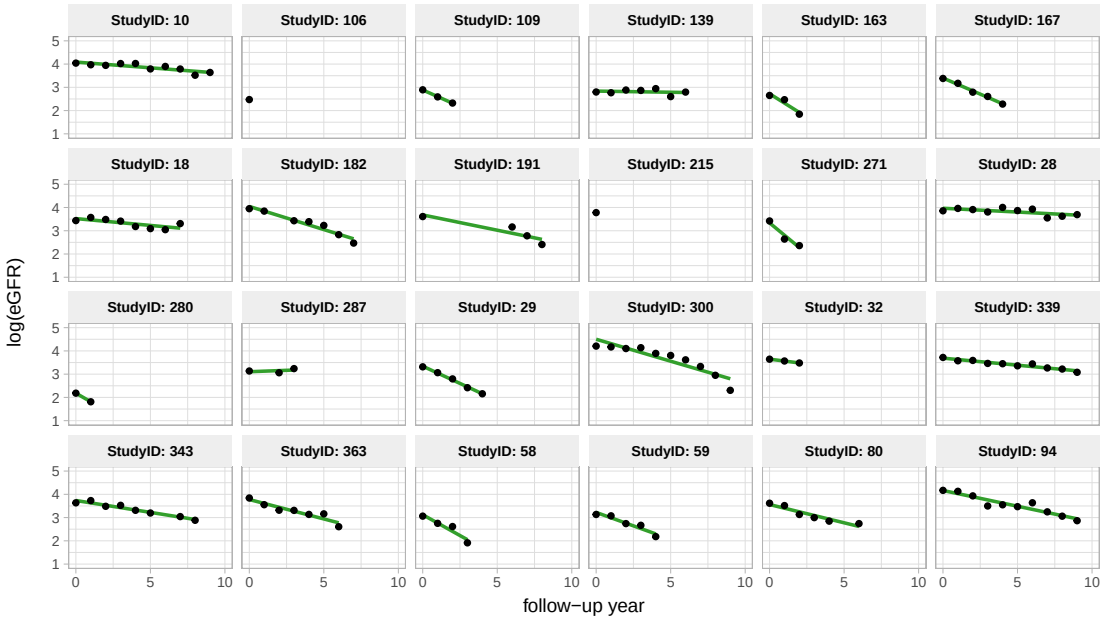

Figure 59: Progression of disease for 24 patients with polycystic kidney disease

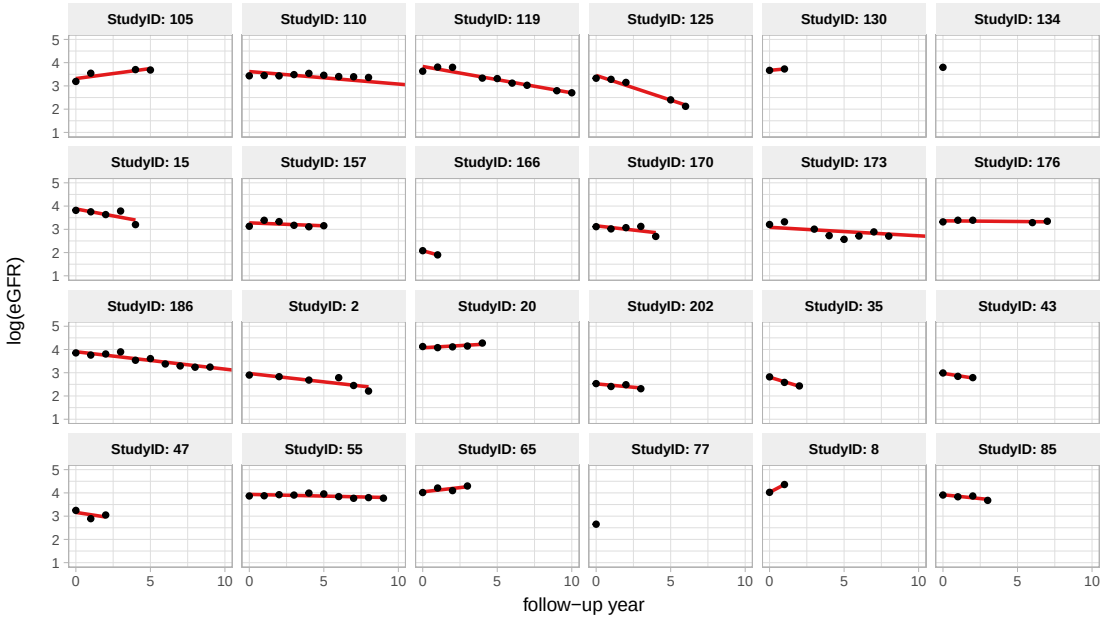

Figure 60: Progression of disease for 24 patients with pyelonephritis

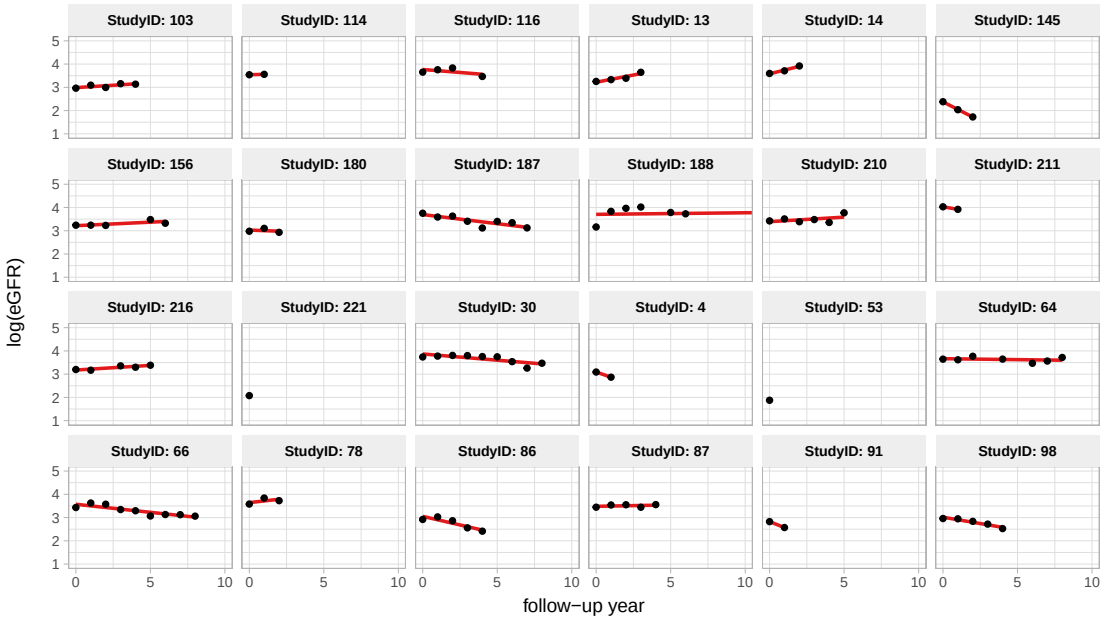

Figure 61: Progression of disease for 24 patients with renovascular disease

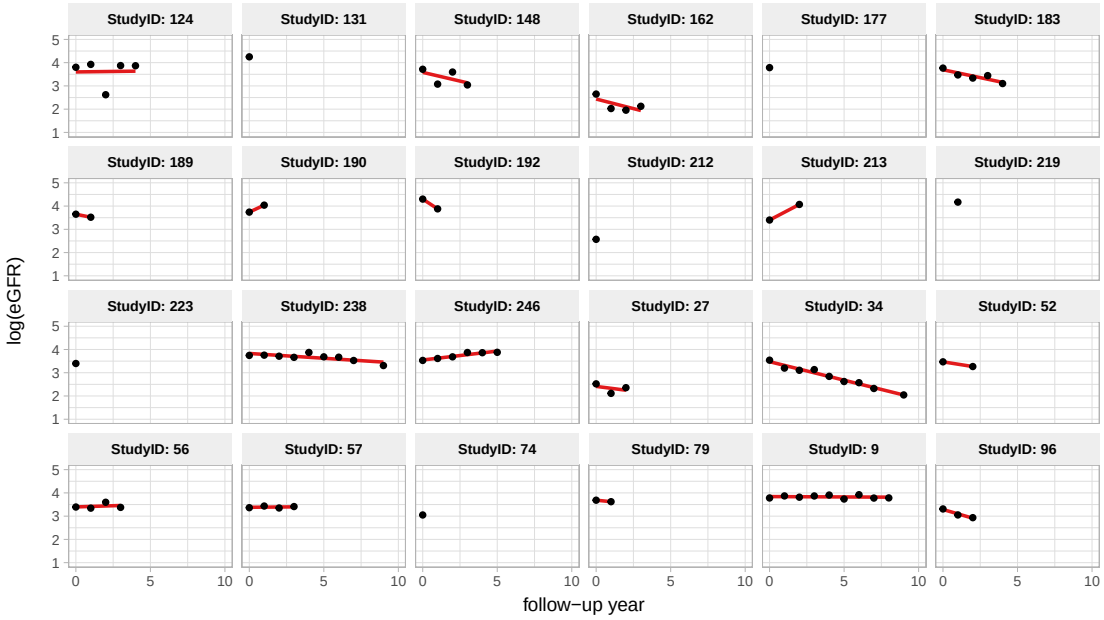

Figure 62: Progression of disease for 24 patients with disease unknown

A.3 Data imputation

We imputed missing values, for each patient, at follow-up appointments, as follows:

- *Continuous variables*: we used the 'Kalman smoothing on a structural model' method to estimate missing values for: BMI, DBP, number of antihypertensives, SBP. Additionally using this method we also imputed the following biochemical markers: CC, CRP, Hb, HbA1c, PO, PTH, Pu, CHO, CO2. This imputation was performed with the na.kalman function from R-package imputeTS (61). If the timeseries had less than 3 measured values then imputation with the function na.kalman was not possible so in these instances we used the spline version of na.interpolation from R-package imputeTS. Before imputation we transformed all the aforementioned continuous variables using the natural logarithm, then after imputation transformed back to the original scale using the exponential function; this ensures all imputed values are positive.
- *Catagoical variables*: we imputed missing values using Last Observation Carried Forward/Backward; when possible Forward was given priority over Backward. This method was applied to comorbidities and weekly alcohol intake. At a given follow-up, if no medications were recorded then this method was used to impute values across all medication categories, otherwise all drugs were assumed to have been recorded leading to the medications' categories being populated as appropriate. Note that EPO treatment is never imputed. We used the na.locf function from R-package imputeTS.

A.4 Dependence between all model variables

A.4.1 Correlation

To assess the correlation between all pairs of explanatory variables we computed a single correlation matrix. This matrix is split into 9 similarly sized sub-matrices which are labelled by (row, column); for example sub-matrix (1,1) is the top left portion of the correlation matrix, similarly sub-matrix (1,2) is the top middle portion and so on. Due to symmetry we only tabulated the upper triangular matrix elements. Tables 31 to 36 show the sub-matrices; correlation values greater than 0.5 are in boldface.

Table 31: Correlation between variables: sub-matrix(1,1)

	age0	bodyMassIndex	CC	comorbidityCancerno	comorbidityCancercurrent	comorbidityCancerprevious	comorbidityCV1	comorbidityCVover 1	comorbidityDiabetestype1	comorbidityDiabetestype2	comorbidityGastrointestinalyes	comorbidityOtheryes	Cr	CRP	DBP	diabetic nephropathy	glomerulonephritis	HKD	obstruction	polycystic kidney
age0	1	-0.1	-0.1	-0.1	0.1	0.1	0	0.3	-0.3	0.3	0.1	0.0	0.1	0.1	-0.2	0.0	-0.2	0.2	0.0	-0.1
bodyMassIndex		1.0	0.0	0.0	0.0	0	0	0.0	0.0	0.2	-0.1	0.0	0.0	0.1	0.1	0.1	0.0	-0.1	0.0	-0.1
CC			1.0	0.0	-0.1	0	0	-0.1	-0.1	0.0	0.0	0.0	-0.1	0.0	0.1	-0.1	0.1	-0.1	0.0	0.0
comorbidityCancerno				1.0	-0.4	**-0.9**	0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.1	0.1	0.0	-0.1	0.0
comorbidityCancercurrent					1.0	-0.1	0	0.1	0.0	0.0	0.1	0.1	0.0	0.2	0.0	0.0	-0.1	0.0	0.0	0.0
comorbidityCancerprevious						1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.1	-0.1
comorbidityCV1							1	-0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	-0.1	0.0	0.0
comorbidityCVover 1								1.0	0.0	0.2	0.1	0.1	0.0	0.0	-0.1	0.1	-0.2	0.1	0.0	-0.1
comorbidityDiabetestype1									1.0	-0.3	-0.1	-0.1	0.1	0.0	0.0	0.3	-0.1	-0.1	0.0	-0.1
comorbidityDiabetestype2										1.0	0.0	0.0	0.1	0.0	-0.2	0.4	-0.2	-0.1	0.0	-0.1
comorbidityGastrointestinalyes											1.0	0.1	-0.1	0.0	0.0	0.0	-0.1	0.0	0.0	0.1
comorbidityOtheryes												1.0	0.0	0.1	0.0	0.0	-0.1	0.0	0.1	0.0
Cr													1.0	0.1	-0.1	0.2	-0.1	0.0	0.0	0.0
CRP														1.0	0.0	0.0	0.0	0.0	0.1	0.0
DBP															1.0	-0.1	0.0	0.0	0.0	0.1
diabetic nephropathy																1.0	-0.3	-0.3	0.0	-0.2
glomerulonephritis																	1.0	-0.1	0.0	-0.1
HKD																		1.0	0.0	-0.1
obstruction																			1.0	0.0
polycystic kidney																				1.0

Table 32: Correlation between variables: sub-matrix(1,2)

	pyelonephritis	renovascular	unknown	ethnicitynonWhite	familyHistoryIHD0yes	followup	followupTime	Hb	HbA1c	livingStatus0alone	logeGFR	med.ACE.ARByes	med.AlphaBlockersyes	med.BetaBlockersyes	med.CCBsyes	med.Diureticsyes	med.Epoyes	med.Ironyes	med.Otheryes	med.ParenteralIronyes
age0	-0.1	0.2	0.1	-0.2	0.0	-0.1	-0.1	-0.1	0.1	0.1	-0.3	-0.1	0.1	0.1	0.1	0.2	0.0	0.1	0.2	0.0
bodyMassIndex	0.1	-0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-0.1	0	0.1	0.0	0.1	0.0	0.2	-0.1	0.0	0.0	0.0
CC	0.0	-0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	-0.1	0.0	-0.1	0.0	0.0	0.0
comorbidityCancerno	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0
comorbidityCancercurrent	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
comorbidityCancerprevious	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0	-0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0
comorbidityCV1	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0
comorbidityCVover 1	0.0	0.2	0.0	0.0	0.1	0.1	0.1	-0.1	0.1	0.0	-0.1	-0.1	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.0
comorbidityDiabetestype1	-0.1	-0.1	-0.1	0.0	0.0	-0.1	0.0	-0.1	0.3	0.0	-0.1	0.1	0.0	-0.1	0.0	0.1	0.1	0.0	0.0	0.1
comorbidityDiabetestype2	-0.1	0.0	-0.1	0.0	0.0	-0.1	-0.1	-0.1	0.4	0.0	-0.2	0.1	0.1	0.1	0.0	0.2	0.1	0.1	0.1	0.1
comorbidityGastrointestinalyes	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.1	-0.1	0.0	0.0	-0.1	-0.1	-0.1	0.0	0.0	0.0
comorbidityOtheryes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	-0.1	-0.1	0.0	0.0	0.1	0.0	0.0
Cr	-0.1	0.0	0.0	-0.1	0.0	0.1	0.1	-0.3	0.0	0.0	**-0.9**	-0.1	0.1	0.0	0.1	0.2	0.3	0.1	0.1	0.2
CRP	0.0	0.0	0.0	0.0	0.0	0.1	0.0	-0.1	0.0	0.0	-0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
DBP	0.1	-0.1	0.0	0.1	0.0	0.0	0.0	0.2	-0.1	0.0	0.2	0.0	-0.1	0.0	-0.1	-0.1	-0.1	-0.1	-0.1	0.0
diabetic nephropathy	-0.2	-0.2	-0.2	0.0	0.0	-0.1	-0.1	-0.2	0.5	0.0	-0.2	0.1	0.1	-0.1	0.1	0.2	0.2	0.1	0.1	0.1
glomerulonephritis	-0.1	-0.1	-0.1	0.0	-0.1	0.1	0.1	0.0	-0.2	0.0	0.1	0.1	0.0	-0.1	0.0	-0.1	0.0	-0.1	0.0	0.0
HKD	-0.1	-0.1	-0.1	0.0	0.0	0.0	0.0	0.1	-0.2	0.0	-0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	-0.1	0.0
obstruction	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
polycystic kidney	-0.1	-0.1	-0.1	0.0	-0.1	0.1	0.1	0.0	-0.1	0.0	0.1	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.1	0.0

Table 33: Correlation between variables: sub-matrix(1,3)

	med.VitaminDyes	numberAKIepisodes	numberAntihypertensives	numberClinicVisits	occupation0ManagerialProfessional	occupation0Intermediate	occupation0NeverWorkedUnemployed	PO	PP	PTH	Pu	SBP	sexfemale	smokingStatus0active	smokingStatus0ex-smoker	totalCholesterol	totalCO2	weeklyAlcohol01 to 14	weeklyAlcohol0over 14
age0	0.1	0.0	0.1	0.0	0.0	0.1	-0.2	0	0.3	0.0	-0.1	0.2	0.0	-0.2	0.2	-0.2	0.0	-0.1	-0.1
bodyMassIndex	0.1	0.0	0.2	0.0	0.0	-0.1	0.1	-0.1	0.0	0.1	0.0	0	0.1	-0.1	0.0	0.0	0.1	0.0	-0.1
CC	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	-0.1	0.0	-0.1	0.0	0.1	0.1	0.0	-0.1	0.1	0.1	0.0	0.0
comorbidityCancerno	0.0	0.0	0.0	0.0	-0.1	0.1	0.0	0	0.0	0.0	0.0	0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
comorbidityCancercurrent	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0	0.0	-0.1	0.1	0.0	0.0	0.0	0.0
comorbidityCancerprevious	0.0	0.0	0.0	0.0	0.1	-0.1	0.0	0	0.0	0.0	0.0	0	0.0	-0.1	0.0	0.0	0.0	-0.1	0.0
comorbidityCV1	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
comorbidityCVover 1	0.1	0.1	0.1	0.0	-0.1	0.0	-0.1	0	0.1	0.0	-0.1	0	-0.1	0.0	0.1	-0.1	0.1	-0.1	-0.1
comorbidityDiabetestype1	0.1	0.0	0.1	0.0	0.1	-0.1	0.1	0.1	0.0	0.1	0.1	0	0.1	0.0	-0.1	0.0	0.0	0.0	0.1
comorbidityDiabetestype2	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.2	0.1	0.0	0.1	-0.1	-0.1	0.1	-0.2	0.0	-0.1	-0.1
comorbidityGastrointestinalyes	0.0	0.1	-0.1	0.0	-0.1	0.1	-0.1	-0.1	0.0	0.0	0.0	0	0.0	0.1	0.0	0.0	0.0	-0.1	0.0
comorbidityOtheryes	0.1	0.1	-0.1	0.0	-0.1	0.0	0.0	0	0.0	0.1	-0.1	0	0.1	0.0	0.1	0.0	0.0	0.0	-0.1
Cr	0.4	0.1	0.1	0.2	0.0	-0.1	0.0	**0.6**	0.1	0.5	0.2	0	-0.1	0.0	0.1	-0.2	-0.3	0.0	0.0
CRP	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DBP	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	0.1	0.4	-0.1	0.0	-0.1	0.2	0.0	0.0	0.0
diabetic nephropathy	0.2	0.1	0.2	0.0	0.1	-0.2	0.0	0.2	0.1	0.2	0.1	0.1	0.0	0.0	0.0	-0.2	0.0	0.0	0.0
glomerulonephritis	-0.1	0.0	0.0	0.1	0.0	0.1	0.0	0	-0.1	-0.1	0.1	-0.1	0.1	0.0	-0.1	0.1	0.0	0.1	0.0
HKD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	-0.1	0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0
obstruction	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
polycystic kidney	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0	-0.1	0.0	-0.1	-0.1	0.0	0.2	-0.1	0.0	0.0	0.0	0.0

Table 34: Correlation between variables: sub-matrix(2,2)

	pyelonephritis	renovascular	unknown	ethnicitynonWhite	familyHistoryIHD0yes	followup	followupTime	Hb	HbA1c	livingStatus0alone	logeGFR	med.ACE.ARByes	med.AlphaBlockersyes	med.BetaBlockersyes	med.CCBsyes	med.Diureticsyes	med.Epoyes	med.Ironyes	med.Otheryes	med.ParenteralIronyes
pyelonephritis	1	-0.1	-0.1	0.1	0.1	0	0	0.1	-0.1	0.0	0.1	0.0	0.0	0.0	-0.1	-0.1	-0.1	0.0	0.0	0.0
renovascular		1.0	-0.1	-0.1	0.1	0	0	0.0	-0.1	0.0	0	0.0	0.1	0.1	0.1	0.0	-0.1	0.0	0.0	-0.1
unknown			1.0	-0.1	0.0	0	0	0.0	-0.1	0.0	0	0.0	-0.1	0.0	-0.1	-0.1	0.0	0.0	0.0	0.0
ethnicitynonWhite				1.0	0.0	0	0	0.0	0.0	-0.1	0.1	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	-0.1	0.0
familyHistoryIHD0yes					1.0	0	0	0.0	0.0	0.0	0	0.0	0.0	0.1	-0.1	0.0	0.0	0.0	0.1	0.0
followup						1	1	0.0	-0.1	0.0	-0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
followupTime							1	0.0	-0.1	0.0	-0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Hb								1.0	-0.1	0.0	0.4	0.0	-0.1	0.0	0.0	-0.2	-0.3	-0.1	-0.2	-0.2
HbA1c									1.0	0.1	-0.1	0.1	0.1	0.0	0.0	0.2	0.0	0.1	0.1	0.0
livingStatus0alone										1.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
logeGFR											1	0.1	-0.1	0.0	-0.1	-0.2	-0.3	-0.1	-0.1	-0.2
med.ACE.ARByes												1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
med.AlphaBlockersyes													1.0	0.0	0.2	0.2	0.1	0.1	0.1	0.0
med.BetaBlockersyes														1.0	0.1	0.1	-0.1	0.0	0.0	0.0
med.CCBsyes															1.0	0.1	0.1	0.1	0.0	0.0
med.Diureticsyes																1.0	0.2	0.1	0.1	0.1
med.Epoyes																	1.0	0.1	0.1	0.3
med.Ironyes																		1.0	0.0	0.0
med.Otheryes																			1.0	0.0
med.ParenteralIronyes																				1.0

Table 35: Correlation between variables: sub-matrix(2,3)

	med.VitaminDyes	numberAKIepisodes	numberAntihypertensives	numberClinicVisits	occupation0ManagerialProfessional	occupation0Intermediate	occupation0NeverWorkedUnemployed	PO	PP	PTH	Pu	SBP	sexfemale	smokingStatus0active	smokingStatus0ex-smoker	totalCholesterol	totalCO2	weeklyAlcohol01 to 14	weeklyAlcohol0over 14
pyelonephritis	0.0	0.0	-0.1	0.0	-0.1	-0.1	0.1	-0.1	-0.1	0.0	0.0	-0.1	0.1	0.0	0.0	0.1	-0.1	-0.1	0.0
renovascular	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0	0.1	0.0	-0.1	0	-0.1	0.1	0.1	0.0	0.0	-0.1	0.0
unknown	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0	0.0	0.0	-0.1	0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0
ethnicitynonWhite	0.0	0.0	0.0	0.0	0.1	0.0	0.1	-0.1	-0.1	0.1	0.0	-0.1	0.0	0.0	0.0	0.1	0.0	-0.1	0.0
familyHistoryIHD0yes	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0	0.0	0.0	0.0	0	0.1	0.0	0.0	0.0	0.0	-0.1	0.0
followup	0.1	0.1	0.0	0.2	0.0	0.0	0.0	0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.1
followupTime	0.1	0.1	0.0	0.2	0.0	0.0	0.0	0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.1
Hb	-0.2	-0.1	-0.1	-0.1	0.0	0.0	0.0	-0.4	-0.2	-0.2	-0.1	0	-0.2	0.0	0.0	0.1	0.1	0.0	0.1
HbA1c	0.1	0.0	0.1	0.0	0.0	-0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	-0.1	-0.1
livingStatus0alone	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.2	0.0	-0.1	0.0	0.0	-0.1	0.0
logeGFR	-0.4	-0.1	-0.2	-0.2	0.0	0.1	0.0	-0.5	-0.2	-0.5	-0.2	-0.1	-0.1	0.0	-0.1	0.1	0.3	0.0	0.1
med.ACE.ARByes	-0.1	-0.1	0.3	0.0	0.0	0.0	0.1	0	0.0	0.0	0.0	-0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1
med.AlphaBlockersyes	0.1	0.0	0.4	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	-0.1	-0.1	0.1	-0.1	0.0	0.1	0.0
med.BetaBlockersyes	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0	0.1	0.1	0.0	0	-0.1	0.0	0.0	-0.1	0.0	0.0	0.1
med.CCBsyes	0.0	-0.1	0.5	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	-0.1	0.0	0.1	-0.1	-0.1	0.1	0.0
med.Diureticsyes	0.1	0.0	0.5	0.1	0.0	0.0	0.0	0.1	0.1	0.2	0.1	0.1	0.1	-0.1	0.1	-0.1	0.1	0.0	-0.1
med.Epoyes	0.2	0.0	0.1	0.2	0.1	0.0	0.0	0.2	0.0	0.2	0.1	0	0.1	-0.1	0.1	-0.1	-0.1	0.0	0.0
med.Ironyes	0.1	0.0	0.1	0.0	0.0	-0.1	0.0	0.1	0.1	0.1	0.0	0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0
med.Otheryes	0.1	0.0	0.1	0.1	0.0	0.0	-0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.1	-0.1	0.0	0.0	0.0
med.ParenteralIronyes	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.1	0.1	0	0.1	0.0	0.0	0.0	-0.1	0.0	0.0

Table 36: Correlation between variables: sub-matrix(3,3)

	med.VitaminDyes	numberAKIepisodes	numberAntihypertensives	numberClinicVisits	occupation0ManagerialProfessional	occupation0Intermediate	occupation0NeverWorkedUnemployed	PO	PP	PTH	Pu	SBP	sexfemale	smokingStatus0active	smokingStatus0ex-smoker	totalCholesterol	totalCO2	weeklyAlcohol01 to 14	weeklyAlcohol0over 14
med.VitaminDyes	1	0.1	0.0	0.1	0.0	-0.1	0.0	0.2	0.1	0.3	0.0	0	0.1	0.0	0.0	-0.1	-0.1	0.0	-0.1
numberAKIepisodes		1.0	-0.1	0.0	0.0	0.0	0.0	0	0.0	0.1	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1
numberAntihypertensives			1.0	0.1	0.1	-0.1	0.0	0.1	0.2	0.1	0.1	0.1	-0.1	-0.1	0.1	-0.2	0.0	0.1	0.1
numberClinicVisits				1.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0
occupation0ManagerialProfessional					1.0	-0.3	-0.1	0	0.0	0.0	0.0	0	-0.2	-0.1	0.1	-0.1	0.0	0.1	0.2
occupation0Intermediate						1.0	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	0.1	0.0	-0.1	0.0	0.0	0.1	0.0
occupation0NeverWorkedUnemployed							1.0	0	0.0	0.0	0.0	0	0.1	0.1	-0.1	0.0	0.0	-0.1	0.0
PO								1	0.1	0.4	0.2	0	0.1	0.1	0.0	0.0	-0.2	0.0	0.0
PP									1.0	0.1	0.2	**0.9**	0.0	0.0	0.1	0.0	0.0	-0.1	0.0
PTH										1.0	0.2	0.1	0.0	0.0	0.0	-0.1	-0.1	-0.1	0.0
Pu											1.0	0.2	0.0	0.1	-0.1	0.2	-0.1	0.0	-0.1
SBP												1	0.0	0.0	0.0	0.1	0.0	-0.1	0.0
sexfemale													1.0	0.0	-0.2	0.2	0.0	-0.1	-0.2
smokingStatus0active														1.0	-0.4	0.1	-0.1	-0.1	0.0
smokingStatus0ex-smoker															1.0	-0.2	0.0	0.0	0.1
totalCholesterol																1.0	0.0	-0.1	0.0
totalCO2																	1.0	0.0	0.0
weeklyAlcohol01 to 14																		1.0	-0.3
weeklyAlcohol0over 14																			1.0

A.4.2 Variance inflation factor

We computed the VIF with the vifstep function from R-package usdm (69). As described in Section 5.1 the function uses a stepwise procedure to exclude highly correlated variables with a VIF above a given threshold. Here we do not include eGFR or Cr.

With a VIF threshold of 5 the method excluded: disease other, SBP, followupTime. With these variables excluded, the VIF values for the remaining variables are given in Table 37.

Reducing the threshold to 2.5 results in the method excluding: disease other, SBP, followupTime, numberAntihypertensives, disease diabetic nephropathy. With these variables excluded, the VIF values for the remaining variables are given in Table 38.

Table 37: Variance inflation factor using all data: threshold 5

Variables	VIF
numberAntihypertensives	4.7
disease diabetic nephropathy	3.5
comorbidityDiabetestype2	2.1
med.Diureticsyes	2.1
disease glomerulonephritis	2.0
age0	2.0
disease HKD	1.9
comorbidityDiabetestype1	1.9
disease renovascular disease	1.8
comorbidityCVover 1	1.8
HbA1c	1.8
med.BetaBlockersyes	1.7
med.CCBsyes	1.7
disease unknown	1.6
med.ACE.ARByes	1.6
med.AlphaBlockersyes	1.6
disease pyelonephritis	1.5
Hb	1.5
PO	1.5
sexfemale	1.5
smokingStatus0active	1.5
smokingStatus0ex-smoker	1.5
disease polycystic kidney disease	1.4
comorbidityCV1	1.4
PTH	1.4
weeklyAlcohol0over 14	1.4
bodyMassIndex	1.3
med.Epoyes	1.3
occupation0ManagerialProfessional	1.3
occupation0Intermediate	1.3
PP	1.3
Pu	1.3
totalCholesterol	1.3
weeklyAlcohol01 to 14	1.3
followup	1.2
CC	1.2
comorbidityGastrointestinalyes	1.2
DBP	1.2

Table 37: Variance inflation factor using all data: threshold 5 *(continued)*

Variables	VIF
ethnicitynonWhite	1.2
med.VitaminDyes	1.2
totalCO2	1.2
disease obstruction	1.1
comorbidityCancercurrent	1.1
comorbidityCancerprevious	1.1
comorbidityOtheryes	1.1
CRP	1.1
familyHistoryIHD0yes	1.1
livingStatus0alone	1.1
med.Ironyes	1.1
med.Otheryes	1.1
med.ParenteralIronyes	1.1
numberAKIepisodes	1.1
numberClinicVisits	1.1
occupation0NeverWorkedUnemployed	1.1

Table 38: Variance inflation factor using all data: threshold 2.5

Variables	VIF
age0	2.0
comorbidityDiabetestype2	2.0
comorbidityCVover 1	1.8
comorbidityDiabetestype1	1.8
HbA1c	1.8
disease glomerulonephritis	1.6
Hb	1.5
PO	1.5
sexfemale	1.5
smokingStatus0active	1.5
smokingStatus0ex-smoker	1.5
disease HKD	1.4
disease renovascular disease	1.4
comorbidityCV1	1.4
med.Diureticsyes	1.4
PTH	1.4
weeklyAlcohol0over 14	1.4
disease polycystic kidney disease	1.3
disease pyelonephritis	1.3
disease unknown	1.3
bodyMassIndex	1.3
med.Epoyes	1.3
occupation0ManagerialProfessional	1.3
occupation0Intermediate	1.3
PP	1.3
Pu	1.3
totalCholesterol	1.3
weeklyAlcohol01 to 14	1.3
followup	1.2
CC	1.2
comorbidityGastrointestinalyes	1.2
DBP	1.2
ethnicitynonWhite	1.2
med.ACE.ARByes	1.2
med.AlphaBlockersyes	1.2
med.BetaBlockersyes	1.2
med.CCBsyes	1.2
med.VitaminDyes	1.2
totalCO2	1.2
disease obstruction	1.1
comorbidityCancercurrent	1.1
comorbidityCancerprevious	1.1
comorbidityOtheryes	1.1
CRP	1.1
familyHistoryIHD0yes	1.1
livingStatus0alone	1.1
med.Ironyes	1.1
med.Otheryes	1.1
med.ParenteralIronyes	1.1
numberAKIepisodes	1.1
numberClinicVisits	1.1
occupation0NeverWorkedUnemployed	1.1

A.5 Linear mixed effects model: residuals by follow-up year

Figures 63-71 show the standardised residual distributions at each follow-up year, they confirm there is no systematic trend over time and that the 95% CI typically covers zero.

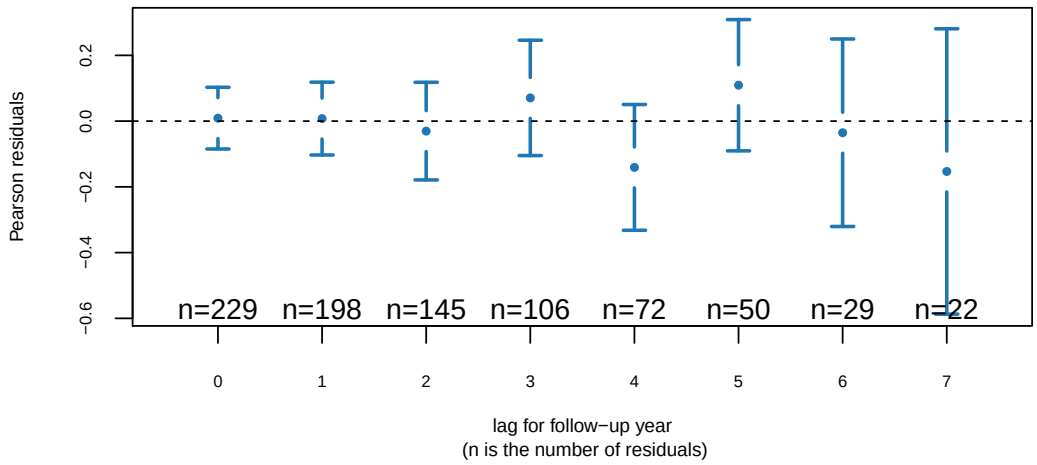

Figure 63: Residuals for diabetic nephropathy model by follow-up year with 95% CIs

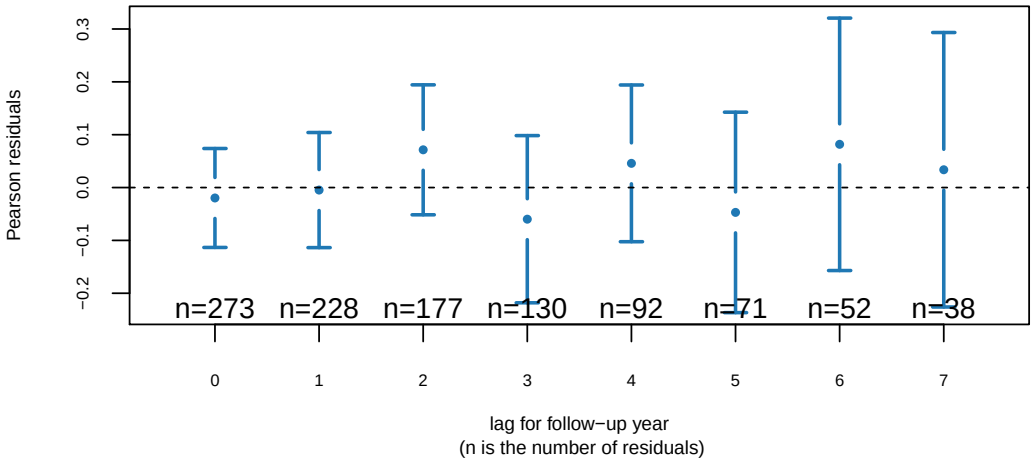

Figure 64: Residuals for glomerulonephritis model by follow-up year with 95% CIs

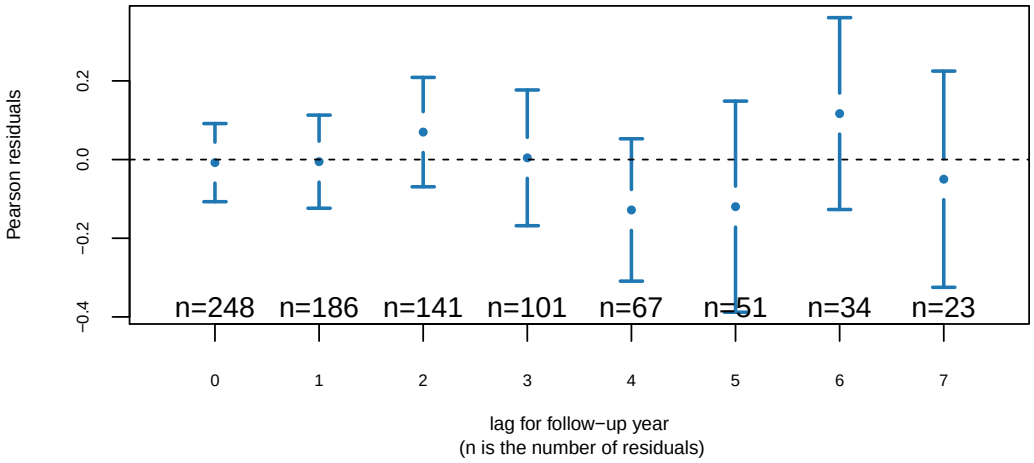

Figure 65: Residuals for HKD model follow-up year with 95% CIs

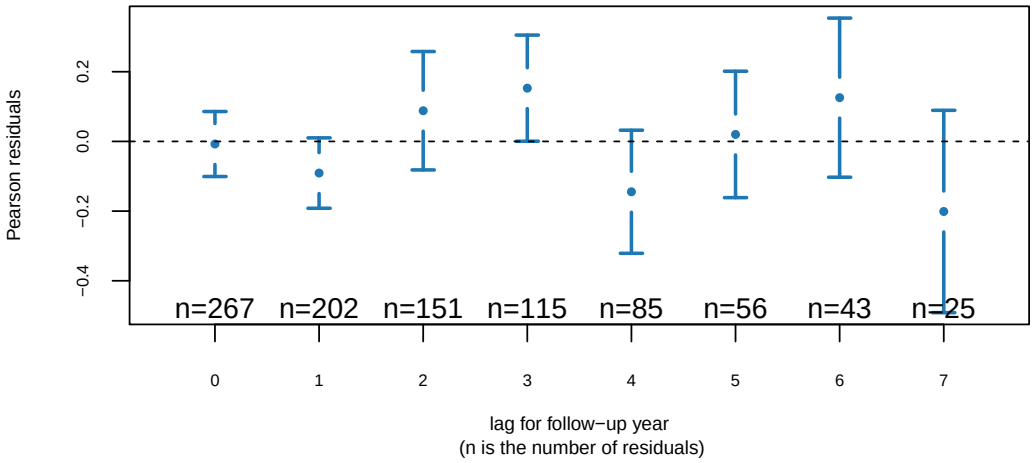

Figure 66: Residuals for disease Other model by follow-up year with 95% CIs

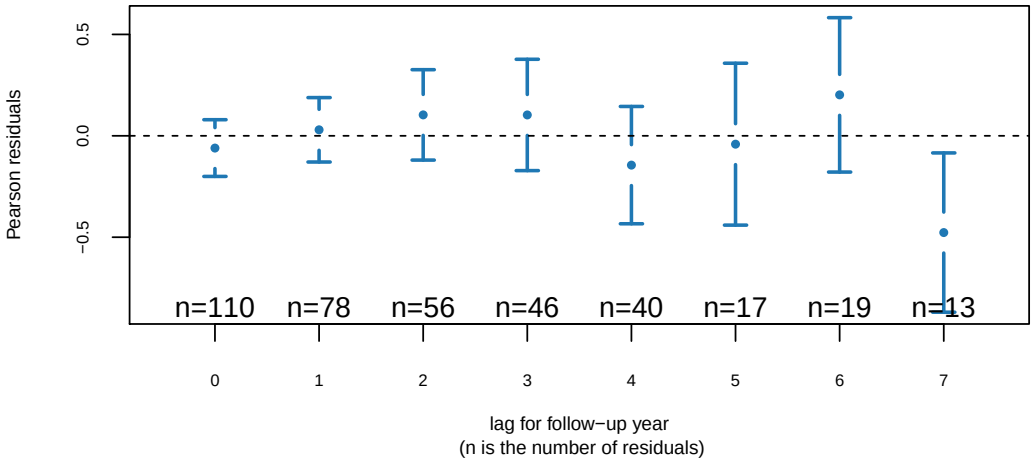

Figure 67: Residuals for polycystic kidney disease model by follow-up year with 95% CIs

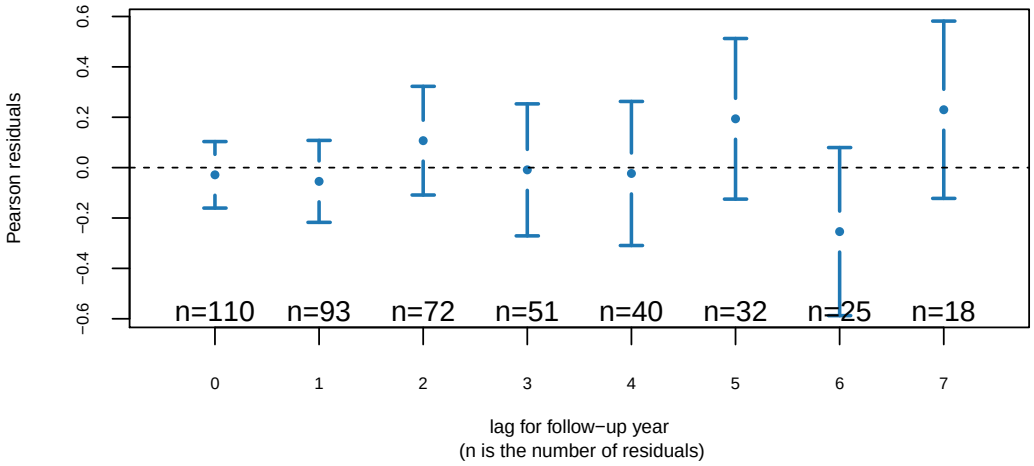

Figure 68: Residuals for pyelonephritis model by follow-up year with 95% CIs

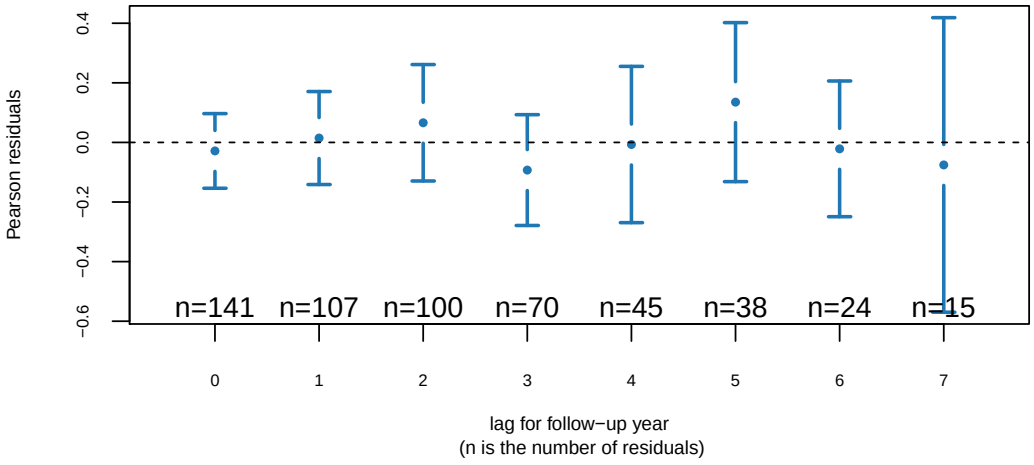

Figure 69: Residuals for renovascular model follow-up year with 95% CIs

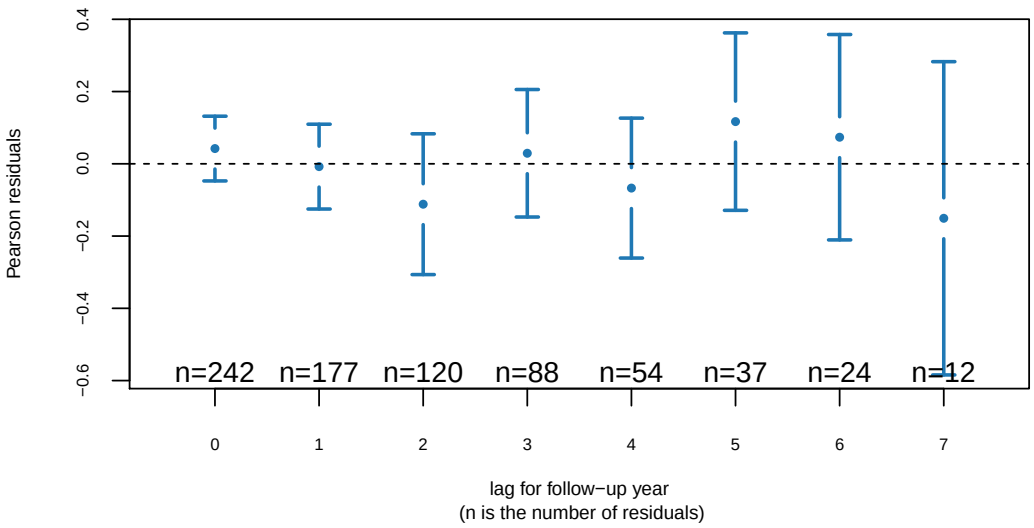

Figure 70: Residuals for unknown disease model follow-up year with 95% CIs

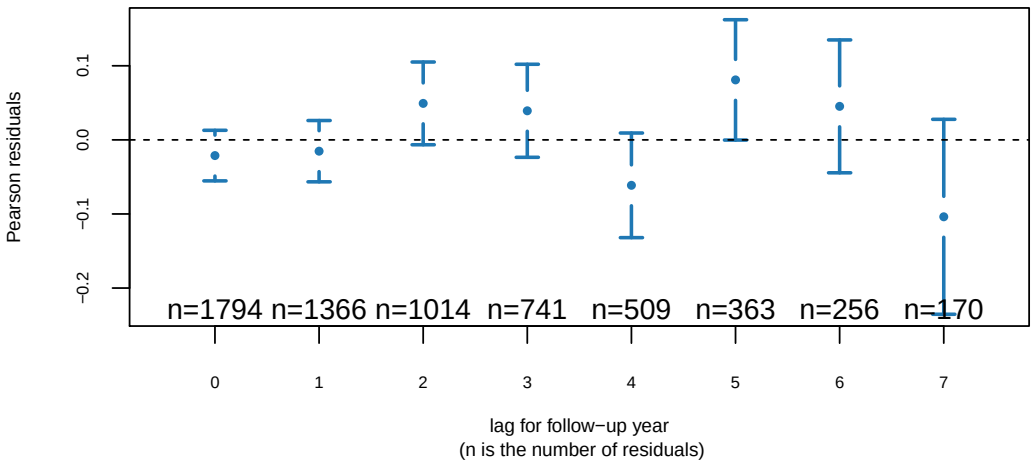

Figure 71: Residuals for single model all diseases by follow-up year with 95% CIs

A.6 Observation counts per factor level for each disease category

The distribution of observations among factor levels for each disease category are tabulated in Tables 39-47; the column labelled *reference* is the factor reference level used in the LME models.

Table 39: Count of observations in each factor level for disease diabetic nephropathy

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	794	13	72	
comorbidityCV	level	no	1	over 1	
	count	237	259	383	
comorbidityDiabetes	level	type2	type1		
	count	714	165		
comorbidityGastrointestinal	level	no	yes		
	count	771	108		
comorbidityOther	level	no	yes		
	count	678	201		
ethnicity	level	White	nonWhite		
	count	840	39		
familyHistoryIHD0	level	no	yes		
	count	490	389		
livingStatus0	level	with others	alone		
	count	734	145		
med.ACE.ARB	level	no	yes		
	count	207	672		
med.AlphaBlockers	level	no	yes		
	count	543	336		
med.BetaBlockers	level	no	yes		
	count	574	305		
med.CCBs	level	no	yes		
	count	420	459		
med.Diuretics	level	no	yes		
	count	236	643		
med.Epo	level	no	yes		
	count	604	275		
med.Iron	level	no	yes		
	count	665	214		
med.Other	level	no	yes		
	count	3	876		
med.ParenteralIron	level	no	yes		
	count	740	139		
med.VitaminD	level	no	yes		
	count	591	288		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	502	308	53	16
sex	level	male	female		
	count	604	275		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	283	101	495	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	491	246	142	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 40: Count of observations in each factor level for disease glomerulonephritis

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	997	16	109	
comorbidityCV	level	no	1	over 1	
	count	699	256	167	
comorbidityDiabetes	level	no	type1	type2	
	count	986	9	127	
comorbidityGastrointestinal	level	no	yes		
	count	1038	84		
comorbidityOther	level	no	yes		
	count	886	236		
ethnicity	level	White	nonWhite		
	count	1080	42		
familyHistoryIHD0	level	no	yes		
	count	679	443		
livingStatus0	level	with others	alone		
	count	931	191		
med.ACE.ARB	level	no	yes		
	count	216	906		
med.AlphaBlockers	level	no	yes		
	count	867	255		
med.BetaBlockers	level	no	yes		
	count	845	277		
med.CCBs	level	no	yes		
	count	636	486		
med.Diuretics	level	no	yes		
	count	649	473		
med.Epo	level	no	yes		
	count	911	211		
med.Iron	level	no	yes		
	count	1011	111		
med.Other	level	no	yes		
	count	56	1066		
med.ParenteralIron	level	no	yes		
	count	1014	108		
med.VitaminD	level	no	yes		
	count	938	184		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	479	367	250	26
sex	level	male	female		
	count	743	379		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	442	116	564	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	412	458	252	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 41: Count of observations in each factor level for disease HKD

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	771	20	90	
comorbidityCV	level	no	1	over 1	
	count	345	200	336	
comorbidityDiabetes	level	no	type2		
	count	727	154		
comorbidityGastrointestinal	level	no	yes		
	count	784	97		
comorbidityOther	level	no	yes		
	count	691	190		
ethnicity	level	White	nonWhite		
	count	846	35		
familyHistoryIHD0	level	no	yes		
	count	497	384		
livingStatus0	level	with others	alone		
	count	691	190		
med.ACE.ARB	level	no	yes		
	count	280	601		
med.AlphaBlockers	level	no	yes		
	count	582	299		
med.BetaBlockers	level	no	yes		
	count	503	378		
med.CCBs	level	no	yes		
	count	350	531		
med.Diuretics	level	no	yes		
	count	365	516		
med.Epo	level	no	yes		
	count	726	155		
med.Iron	level	no	yes		
	count	732	149		
med.Other	level	no	yes		
	count	26	855		
med.ParenteralIron	level	no	yes		
	count	795	86		
med.VitaminD	level	no	yes		
	count	655	226		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	424	263	171	23
sex	level	male	female		
	count	580	301		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	333	64	484	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	429	301	151	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 42: Count of observations in each factor level for disease other

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	733	62	187	
comorbidityCV	level	no	1	over 1	
	count	459	284	239	
comorbidityDiabetes	level	no	type1	type2	
	count	776	5	201	
comorbidityGastrointestinal	level	no	yes		
	count	882	100		
comorbidityOther	level	no	yes		
	count	747	235		
ethnicity	level	White	nonWhite		
	count	956	26		
familyHistoryIHD0	level	no	yes		
	count	542	440		
livingStatus0	level	with others	alone		
	count	868	114		
med.ACE.ARB	level	no	yes		
	count	442	540		
med.AlphaBlockers	level	no	yes		
	count	789	193		
med.BetaBlockers	level	no	yes		
	count	714	268		
med.CCBs	level	no	yes		
	count	637	345		
med.Diuretics	level	no	yes		
	count	655	327		
med.Epo	level	no	yes		
	count	840	142		
med.Iron	level	no	yes		
	count	855	127		
med.Other	level	no	yes		
	count	36	946		
med.ParenteralIron	level	no	yes		
	count	910	72		
med.VitaminD	level	no	yes		
	count	802	180		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	484	280	172	46
sex	level	male	female		
	count	606	376		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	371	134	477	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	486	320	176	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 43: Count of observations in each factor level for disease PKD

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	383	5	14	
comorbidityCV	level	no	1	over 1	
	count	242	102	58	
comorbidityDiabetes	level	no	type2		
	count	374	28		
comorbidityGastrointestinal	level	no	yes		
	count	307	95		
comorbidityOther	level	no	yes		
	count	299	103		
ethnicity	level	White	nonWhite		
	count	399	3		
familyHistoryIHD0	level	no	yes		
	count	247	155		
livingStatus0	level	with others	alone		
	count	337	65		
med.ACE.ARB	level	no	yes		
	count	98	304		
med.AlphaBlockers	level	no	yes		
	count	304	98		
med.BetaBlockers	level	no	yes		
	count	286	116		
med.CCBs	level	no	yes		
	count	230	172		
med.Diuretics	level	no	yes		
	count	255	147		
med.Epo	level	no	yes		
	count	363	39		
med.Iron	level	no	yes		
	count	358	44		
med.Other	level	no	yes		
	count	46	356		
med.ParenteralIron	level	no	yes		
	count	373	29		
med.VitaminD	level	no	yes		
	count	336	66		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	153	150	75	24
sex	level	male	female		
	count	202	200		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	160	49	193	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	162	136	104	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 44: Count of observations in each factor level for disease pyelonephritis

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	416	10	34	
comorbidityCV	level	no	1	over 1	
	count	258	114	88	
comorbidityDiabetes	level	no	type2		
	count	397	63		
comorbidityGastrointestinal	level	no	yes		
	count	425	35		
comorbidityOther	level	no	yes		
	count	360	100		
ethnicity	level	White	nonWhite		
	count	447	13		
familyHistoryIHD0	level	no	yes		
	count	235	225		
livingStatus0	level	with others	alone		
	count	370	90		
med.ACE.ARB	level	no	yes		
	count	166	294		
med.AlphaBlockers	level	no	yes		
	count	369	91		
med.BetaBlockers	level	no	yes		
	count	351	109		
med.CCBs	level	no	yes		
	count	299	161		
med.Diuretics	level	no	yes		
	count	310	150		
med.Epo	level	no	yes		
	count	424	36		
med.Iron	level	no	yes		
	count	413	47		
med.Other	level	no	yes		
	count	35	425		
med.ParenteralIron	level	no	yes		
	count	430	30		
med.VitaminD	level	no	yes		
	count	345	115		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	211	116	111	22
sex	level	male	female		
	count	216	244		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	200	63	197	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	242	143	75	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 45: Count of observations in each factor level for disease renovascular

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	472	25	65	
comorbidityCV	level	no	1	over 1	
	count	82	100	380	
comorbidityDiabetes	level	no	type1	type2	
	count	382	3	177	
comorbidityGastrointestinal	level	no	yes		
	count	487	75		
comorbidityOther	level	no	yes		
	count	419	143		
ethnicity	level	White	nonWhite		
	count	558	4		
familyHistoryIHD0	level	no	yes		
	count	269	293		
livingStatus0	level	with others	alone		
	count	437	125		
med.ACE.ARB	level	no	yes		
	count	203	359		
med.AlphaBlockers	level	no	yes		
	count	308	254		
med.BetaBlockers	level	no	yes		
	count	287	275		
med.CCBs	level	no	yes		
	count	201	361		
med.Diuretics	level	no	yes		
	count	172	390		
med.Epo	level	no	yes		
	count	491	71		
med.Iron	level	no	yes		
	count	458	104		
med.ParenteralIron	level	no	yes		
	count	526	36		
med.VitaminD	level	no	yes		
	count	445	117		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	337	130	90	5
sex	level	male	female		
	count	371	191		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	84	101	377	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	305	147	110	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 46: Count of observations in each factor level for disease unknown

factors		reference			
comorbidityCancer	level	no	current	previous	
	count	677	14	75	
comorbidityCV	level	no	1	over 1	
	count	304	190	272	
comorbidityDiabetes	level	no	type1	type2	
	count	615	2	149	
comorbidityGastrointestinal	level	no	yes		
	count	633	133		
comorbidityOther	level	no	yes		
	count	561	205		
ethnicity	level	White	nonWhite		
	count	745	21		
familyHistoryIHD0	level	no	yes		
	count	406	360		
livingStatus0	level	with others	alone		
	count	559	207		
med.ACE.ARB	level	no	yes		
	count	276	490		
med.AlphaBlockers	level	no	yes		
	count	571	195		
med.BetaBlockers	level	no	yes		
	count	480	286		
med.CCBs	level	no	yes		
	count	439	327		
med.Diuretics	level	no	yes		
	count	389	377		
med.Epo	level	no	yes		
	count	626	140		
med.Iron	level	no	yes		
	count	621	145		
med.Other	level	no	yes		
	count	33	733		
med.ParenteralIron	level	no	yes		
	count	682	84		
med.VitaminD	level	no	yes		
	count	608	158		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	363	229	147	27
sex	level	male	female		
	count	441	325		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	262	69	435	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	453	209	104	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

Table 47: Count of observations in each factor level for single model all diseases

factors		reference								
comorbidityCancer	level	no	current	previous						
	count	5563	183	701						
comorbidityCV	level	no	1	over 1						
	count	2752	1600	2095						
comorbidityDiabetes	level	no	type1	type2						
	count	4320	221	1906						
comorbidityGastrointestinal	level	no	yes							
	count	5677	770							
comorbidityOther	level	no	yes							
	count	4931	1516							
disease[†]	level	Ot	DN	GN	Ob	PKD	PN	RVD	HKD	Un
	count	982	1223	1122	49	402	460	562	881	766
ethnicity	level	White	nonWhite							
	count	6254	193							
familyHistoryIHD0	level	no	yes							
	count	3563	2884							
livingStatus0	level	with others	alone							
	count	5247	1200							
med.ACE.ARB	level	no	yes							
	count	2006	4441							
med.AlphaBlockers	level	no	yes							
	count	4596	1851							
med.BetaBlockers	level	no	yes							
	count	4278	2169							
med.CCBs	level	no	yes							
	count	3406	3041							
med.Diuretics	level	no	yes							
	count	3168	3279							
med.Epo	level	no	yes							
	count	5274	1173							

Table 47: Count of observations in each factor level for single model all diseases
(continued)

factors		reference			
med.Iron	level	no	yes		
	count	5413	1034		
med.Other	level	no	yes		
	count	236	6211		
med.ParenteralIron	level	no	yes		
	count	5788	659		
med.VitaminD	level	no	yes		
	count	4998	1449		
occupation0*	level	RoutMan	ManaProf	Interm	Unempl
	count	3150	1961	1139	197
sex	level	male	female		
	count	4014	2433		
smokingStatus0	level	non-smoker	active	ex-smoker	
	count	2247	743	3457	
weeklyAlcohol0	level	under 1	1 to 14	over 14	
	count	3211	2065	1171	

* abbreviations: RoutMan=RoutineManual, ManaProf=ManagerialProfessional, Interm=Intermediate, Unempl=NeverWorkedUnemployed

† abbreviations: DN=diabetic nephropathy, GN=glomerulonephritis, Ob=obstruction, Ot=other, PKD=polycystic kidney disease, PN=pyelonephritis, RVD=renovascular, HKD=hypertensive kidney disease, Un=unknown

A.7 Unstandardised model

We now consider the unstandardised model fixed effect regression parameters for our disease categories. Unlike the standardised model we work in the original units of measure for each explanatory variable i.e. there is no re-scaling. For the sake of comparability with the standardised results we set the step change in each parameter equal to one standard deviation of the corresponding explanatory variable i.e. $\theta_r = \sigma_r$ relates to $\theta_r^{'} = 1$ since $\theta_r^{'} = \theta_r/\sigma_r$ (Section 4.5 gives more details). The advantage here is that we do not make an arbitrary rescaling of any regression parameters and therefore the results are easier to interpret, particularly in relation to rates of change with respect to time (see Section 4.4). As discussed in Section 4.5 we report results in terms of the *relative change* in eGFR induced by a step change in the parameter of interest. As previously p-values are reported at the 0.05 significance level.

For each disease model the details of the fixed effect parameter estimates are given in Tables 48 to 57. Additionally Figures 72 to 80 summarise the relative change in eGFR for $\theta_r = \sigma_r$ and also indicate the clinically significant level of a 5% change in eGFR.

A.7.1 Overview

Table 48: model summary for each disease

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
(Intercept)	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$	$(+)^{***}$
age0	$(-)^{**}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{**}$	$(-)^{*}$	$(-)^{***}$	$(-)^{***}$
bodyMassIndex	$(-)^{**}$					$(+)^{\sim}$			
bodyMassIndex:followupTime	$(+)^{\sim}$					$(+)^{\sim}$			
CC	$(-)^{**}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{**}$
CC:followupTime	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$
comorbidityCancercurrent			$(+)^{\sim}$			$(-)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$
comorbidityCancercurrent:followupTime			$(+)^{\sim}$			$(+)^{\sim}$		$(+)^{\sim}$	$(-)^{\sim}$
comorbidityCancerprevious			$(-)^{*}$			$(-)^{\sim}$		$(+)^{\sim}$	$(-)^{\sim}$
comorbidityCancerprevious:followupTime			$(+)^{\sim}$			$(+)^{\sim}$		$(-)^{\sim}$	$(+)^{\sim}$
comorbidityCV1			$(-)^{*}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{*}$
comorbidityCV1:followupTime			$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$
comorbidityCVover 1			$(-)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{*}$		$(-)^{**}$
comorbidityCVover 1:followupTime			$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$
comorbidityDiabetestype1				$(-)^{\sim}$			$(-)^{\sim}$		
comorbidityDiabetestype1:followupTime				$(-)^{\sim}$			$(+)^{\sim}$		
comorbidityDiabetestype2				$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$		
comorbidityDiabetestype2:followupTime				$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$		
comorbidityGastrointestinal						$(+)^{\sim}$	$(+)^{\sim}$		
comorbidityGastrointestinal:followupTime						$(+)^{\sim}$	$(-)^{\sim}$		
comorbidityOther			$(+)^{\sim}$						
comorbidityOther:followupTime			$(-)^{\sim}$						
CRP		$(+)^{\sim}$		$(+)^{\sim}$	$(-)^{\sim}$		$(+)^{\sim}$		
CRP:followupTime		$(+)^{\sim}$		$(-)^{\sim}$	$(+)^{\sim}$		$(-)^{\sim}$		

Table 48: model summary for each disease *(continued)*

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
DBP	$(+)^{*}$	$(-)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{*}$	$(+)^{*}$
DBP:followupTime	$(-)^{\sim}$	$(+)^{*}$	$(+)^{\sim}$		$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{*}$	$(-)^{\sim}$	$(+)^{\sim}$
disease diabetic nephropathy									$(-)^{**}$
disease glomerulonephritis									$(+)^{\sim}$
disease HKD									$(-)^{\sim}$
disease obstruction									$(-)^{***}$
disease polycystic kidney disease									$(-)^{***}$
disease pyelonephritis									$(-)^{**}$
disease renovascular disease									$(-)^{\sim}$
disease unknown									$(-)^{\sim}$
ethnicitynonWhite			$(+)^{\sim}$						
familyHistoryIHD0					$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$		
followupTime	$(-)^{**}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$
Hb	$(+)^{\sim}$	$(+)^{**}$	$(+)^{***}$	$(+)^{***}$	$(+)^{\sim}$	$(+)^{*}$	$(+)^{**}$	$(+)^{***}$	$(+)^{***}$
Hb:followupTime	$(+)^{**}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{*}$	$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{**}$
med.ACE.ARB	$(+)^{***}$	$(+)^{**}$	$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{**}$
med.ACE.ARB:followupTime	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$
med.AlphaBlockers		$(-)^{\sim}$				$(-)^{\sim}$	$(+)^{\sim}$		$(-)^{\sim}$
med.AlphaBlockers:followupTime		$(-)^{\sim}$				$(+)^{\sim}$	$(+)^{**}$		$(-)^{\sim}$
med.BetaBlockers	$(+)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$				$(+)^{\sim}$	
med.BetaBlockers:followupTime	$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$				$(-)^{\sim}$	
med.CCBs		$(-)^{**}$	$(-)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$	$(-)^{*}$
med.CCBs:followupTime		$(+)^{\sim}$	$(+)^{\sim}$		$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$	$(-)^{\sim}$
med.Diuretics					$(+)^{\sim}$		$(-)^{*}$		$(-)^{**}$

Table 48: model summary for each disease *(continued)*

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
med.Diuretics:followupTime					$(-)^{\sim}$		$(+)^{\sim}$		$(+)^{\sim}$
med.Epo	$(-)^{**}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{*}$	$(-)^{***}$
med.Epo:followupTime	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$
med.Iron	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$
med.Iron:followupTime	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{*}$
med.Other									$(-)^{*}$
med.Other:followupTime									$(+)^{\sim}$
med.ParenteralIron	$(-)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$	$(+)^{\sim}$	$(+)^{*}$	$(-)^{\sim}$	$(-)^{\sim}$
med.ParenteralIron:followupTime	$(+)^{**}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$
med.VitaminD	$(-)^{***}$	$(-)^{***}$	$(-)^{**}$	$(-)^{***}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{*}$	$(-)^{**}$	$(-)^{***}$
med.VitaminD:followupTime	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$
numberAKIepisodes	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$		$(-)^{\sim}$		$(+)^{*}$	$(-)^{\sim}$
numberAKIepisodes:followupTime	$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$		$(+)^{\sim}$		$(-)^{\sim}$	$(-)^{\sim}$
numberAntihypertensives	$(-)^{*}$	$(-)^{\sim}$	$(-)^{**}$		$(-)^{\sim}$	$(-)^{**}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$
numberAntihypertensives:followupTime	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{*}$	$(-)^{\sim}$
numberClinicVisits			$(-)^{\sim}$	$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$			$(-)^{\sim}$
numberClinicVisits:followupTime			$(-)^{***}$	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$			$(-)^{\sim}$
occupation0ManagerialProfessional					$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$		
occupation0Intermediate					$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$		
occupation0NeverWorkedUnemployed					$(+)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$		
PO	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{*}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$
PO:followupTime	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{***}$	$(-)^{**}$
PP	$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$		$(+)^{*}$	$(-)^{\sim}$	$(+)^{**}$	$(+)^{\sim}$	$(+)^{*}$
PP:followupTime	$(-)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$		$(+)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$

Table 48: model summary for each disease *(continued)*

parameter	diabetic neph.	glomerulonephritis	HKD	other	PKD	pyelonephritis	renovascular	unknown	all[1]
PTH	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$	$(-)^{**}$	$(-)^{\sim}$	$(-)^{**}$	$(-)^{***}$	$(-)^{***}$
PTH:followupTime	$(+)^{\sim}$	$(+)^{***}$	$(-)^{\sim}$	$(+)^{**}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{***}$	$(+)^{***}$
Pu	$(+)^{\sim}$	$(+)^{**}$	$(-)^{\sim}$		$(-)^{*}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$	$(+)^{\sim}$
Pu:followupTime	$(-)^{***}$	$(-)^{***}$	$(-)^{***}$		$(+)^{\sim}$	$(-)^{***}$	$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{***}$
sexfemale						$(-)^{\sim}$			
smokingStatus0active			$(-)^{\sim}$			$(-)^{\sim}$	$(-)^{\sim}$		
smokingStatus0ex-smoker			$(-)^{\sim}$			$(+)^{\sim}$	$(-)^{\sim}$		
totalCholesterol	$(+)^{\sim}$			$(+)^{\sim}$			$(+)^{\sim}$	$(+)^{\sim}$	$(+)^{*}$
totalCholesterol:followupTime	$(-)^{\sim}$			$(-)^{\sim}$			$(-)^{**}$	$(+)^{\sim}$	$(-)^{\sim}$
totalCO2		$(+)^{***}$	$(+)^{**}$	$(+)^{***}$	$(+)^{*}$	$(+)^{**}$		$(+)^{**}$	$(+)^{***}$
totalCO2:followupTime		$(-)^{\sim}$	$(-)^{\sim}$	$(+)^{\sim}$	$(-)^{\sim}$	$(-)^{*}$		$(-)^{\sim}$	$(-)^{***}$
weeklyAlcohol01 to 14			$(-)^{*}$			$(-)^{\sim}$	$(-)^{\sim}$	$(-)^{\sim}$	
weeklyAlcohol0over 14			$(-)^{\sim}$			$(+)^{*}$	$(+)^{\sim}$	$(+)^{\sim}$	

Note:

regression parameter sign: positive (+); negative (-)

p-value significance levels: <0.001 ***; 0.001-0.01 **; 0.01-0.05 *; >0.05 $\sim$

[1] 'all' denotes 'single model all diseases'

A.7.2 Diabetic nephropathy

Table 49: Estimated changes in outcome for changes in parameters for disease diabetic nephropathy

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\frac{\mathbb{E}(\Delta^r \hat{Y}^*)}{\mathbb{E}(\hat{Y}^*_{ij}) = 30}$
biochemical	CC	-0.2360	8.4e-02	0.1	-3.42	-1.03
	CC:followupTime	0.0124	2.4e-02	5.6	7.15	2.15
	Hb	0.0010	8.6e-04	17.1	1.64	0.49
	Hb:followupTime	0.0009	3.0e-04	304.5	32.22	9.67
	PO	-0.3684	5.3e-02	0.3	-9.93	-2.98
	PO:followupTime	0.0210	1.5e-02	2.9	6.27	1.88
	PTH	-0.0060	1.2e-03	11.8	-6.82	-2.05
	PTH:followupTime	0.0002	2.2e-04	52.3	0.83	0.25
	Pu	0.0134	8.2e-03	1.6	2.10	0.63
	Pu:followupTime	-0.0091	2.5e-03	5.3	-4.69	-1.41
	totalCholesterol	0.0181	1.3e-02	1.1	2.10	0.63
	totalCholesterol:followupTime	-0.0005	4.5e-03	11.0	-0.56	-0.17
catagorical	med.ACE.ARB	0.1268	3.4e-02	1.0	13.51	4.05
	med.ACE.ARB:followupTime	-0.0178	1.1e-02	1.0	-1.76	-0.53
	med.BetaBlockers	0.0583	3.3e-02	1.0	6.01	1.80
	med.BetaBlockers:followupTime	-0.0073	9.9e-03	1.0	-0.73	-0.22
	med.Epo	-0.0865	2.8e-02	1.0	-8.28	-2.48
	med.Epo:followupTime	0.0082	8.4e-03	1.0	0.82	0.25
	med.Iron	-0.0285	3.1e-02	1.0	-2.81	-0.84
	med.Iron:followupTime	0.0078	1.1e-02	1.0	0.79	0.24
	med.ParenteralIron	-0.0518	2.6e-02	1.0	-5.04	-1.51
	med.ParenteralIron:followupTime	0.0260	9.0e-03	1.0	2.64	0.79
	med.VitaminD	-0.1334	3.0e-02	1.0	-12.49	-3.75
	med.VitaminD:followupTime	0.0035	8.5e-03	1.0	0.35	0.10
general	age0	-0.0057	1.8e-03	1.0	-0.56	-0.17
	bodyMassIndex	-0.0099	2.9e-03	5.9	-5.67	-1.70
	bodyMassIndex:followupTime	0.0015	9.0e-04	69.8	10.86	3.26
	DBP	0.0021	1.0e-03	11.3	2.44	0.73
	DBP:followupTime	-0.0001	3.4e-04	178.3	-2.22	-0.67
	followupTime	-0.2179	7.9e-02	1.0	-19.58	-5.87
	numberAKIepisodes	-0.0324	2.6e-02	0.4	-1.14	-0.34
	numberAKIepisodes:followupTime	0.0034	5.8e-03	1.4	0.50	0.15
	numberAntihypertensives	-0.0277	1.2e-02	1.4	-3.84	-1.15
	numberAntihypertensives:followupTime	-0.0017	3.6e-03	7.5	-1.29	-0.39
	PP	0.0014	5.8e-04	19.3	2.83	0.85
	PP:followupTime	-0.0001	2.1e-04	171.0	-1.21	-0.36

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

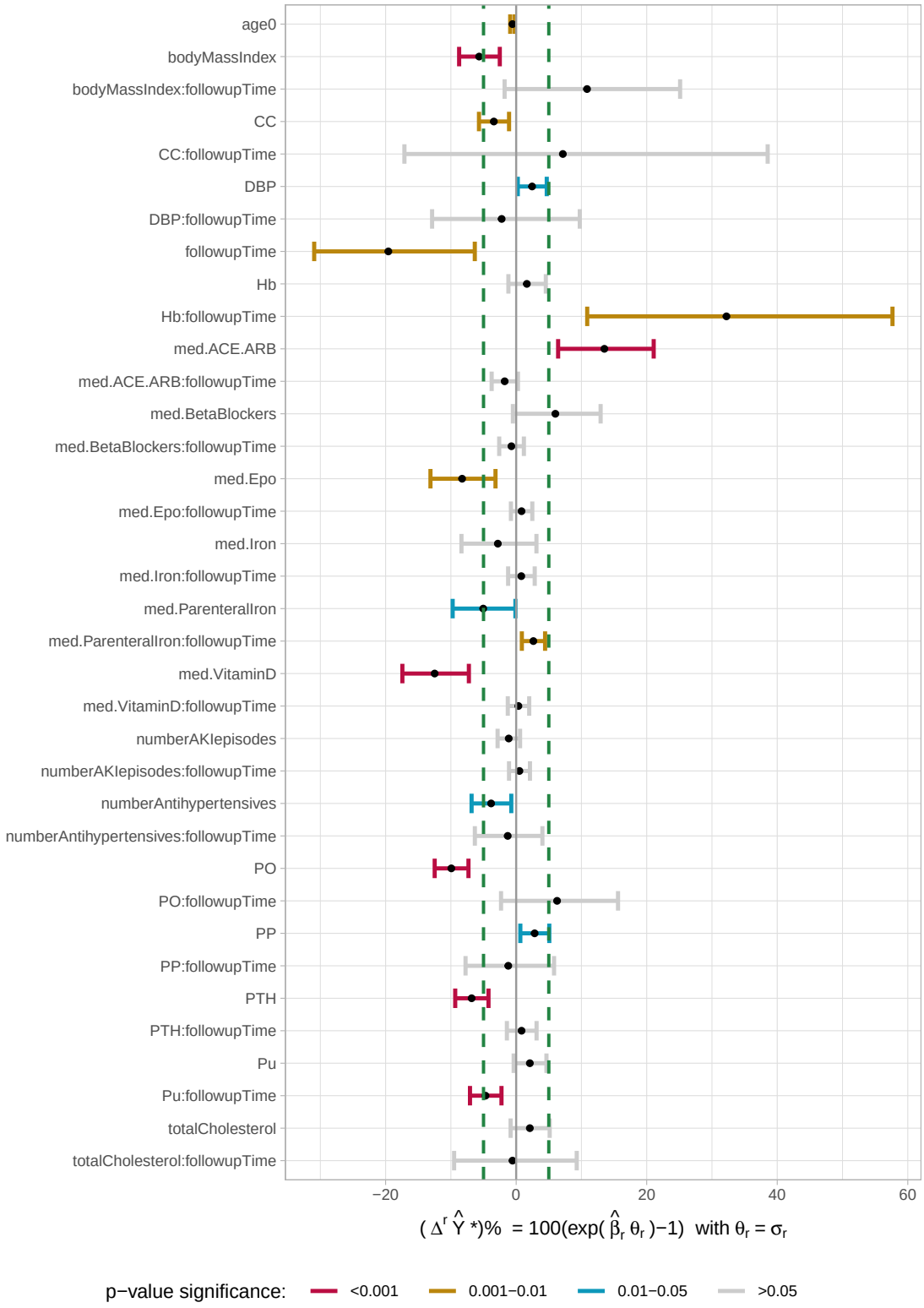

Figure 72: Relative change in eGFR for un-standardised model using 95% CIs: diabetic nephropathy

A.7.3 Glomerulonephritis

Table 50: Estimated changes in outcome for changes in parameters for disease glomerulonephritis

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\frac{\mathbb{E}(\Delta^r \hat{Y}^*)}{\mathbb{E}(\hat{Y}^*_{ij}) = 30}$
biochemical	CC	0.0523	9.0e-02	0.1	0.77	0.23
	CC:followupTime	-0.0135	2.2e-02	5.6	-7.28	-2.18
	CRP	0.0018	1.3e-03	19.9	3.55	1.07
	CRP:followupTime	0.0001	3.2e-04	71.0	0.50	0.15
	Hb	0.0027	8.1e-04	17.1	4.69	1.41
	Hb:followupTime	0.0002	2.3e-04	304.5	7.85	2.36
	PO	-0.4846	5.0e-02	0.3	-12.86	-3.86
	PO:followupTime	-0.0058	1.4e-02	2.9	-1.67	-0.50
	PTH	-0.0091	1.7e-03	11.8	-10.20	-3.06
	PTH:followupTime	0.0017	4.7e-04	52.3	9.36	2.81
	Pu	0.0190	6.3e-03	1.6	3.00	0.90
	Pu:followupTime	-0.0081	2.0e-03	5.3	-4.19	-1.26
	totalCO2	0.0154	3.8e-03	3.5	5.61	1.68
	totalCO2:followupTime	-0.0013	1.0e-03	56.0	-6.94	-2.08
catagorical	med.ACE.ARB	0.1015	3.6e-02	1.0	10.69	3.21
	med.ACE.ARB:followupTime	-0.0137	9.2e-03	1.0	-1.37	-0.41
	med.AlphaBlockers	-0.0294	3.8e-02	1.0	-2.90	-0.87
	med.AlphaBlockers:followupTime	-0.0052	1.0e-02	1.0	-0.52	-0.16
	med.BetaBlockers	-0.0569	3.6e-02	1.0	-5.53	-1.66
	med.BetaBlockers:followupTime	-0.0019	1.1e-02	1.0	-0.19	-0.06
	med.CCBs	-0.1106	3.4e-02	1.0	-10.47	-3.14
	med.CCBs:followupTime	0.0002	8.8e-03	1.0	0.02	0.01
	med.Epo	-0.0417	3.3e-02	1.0	-4.09	-1.23
	med.Epo:followupTime	-0.0064	8.4e-03	1.0	-0.63	-0.19
	med.Iron	0.0350	4.1e-02	1.0	3.56	1.07
	med.Iron:followupTime	-0.0173	1.2e-02	1.0	-1.71	-0.51
	med.ParenteralIron	-0.0136	3.4e-02	1.0	-1.35	-0.41
	med.ParenteralIron:followupTime	-0.0062	9.6e-03	1.0	-0.62	-0.19
	med.VitaminD	-0.1598	3.8e-02	1.0	-14.77	-4.43
	med.VitaminD:followupTime	-0.0112	1.0e-02	1.0	-1.11	-0.33
general	age0	-0.0060	1.7e-03	1.0	-0.60	-0.18
	DBP	-0.0015	1.1e-03	11.3	-1.70	-0.51
	DBP:followupTime	0.0007	2.9e-04	178.3	12.36	3.71
	followupTime	-0.0286	6.7e-02	1.0	-2.82	-0.85
	numberAKIepisodes	-0.0269	4.7e-02	0.4	-0.94	-0.28
	numberAKIepisodes:followupTime	0.0077	1.4e-02	1.4	1.12	0.34
	numberAntihypertensives	-0.0040	1.6e-02	1.4	-0.56	-0.17
	numberAntihypertensives:followupTime	-0.0004	3.9e-03	7.5	-0.32	-0.10
	PP	0.0008	7.1e-04	19.3	1.57	0.47
	PP:followupTime	0.0000	2.0e-04	171.0	0.10	0.03

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

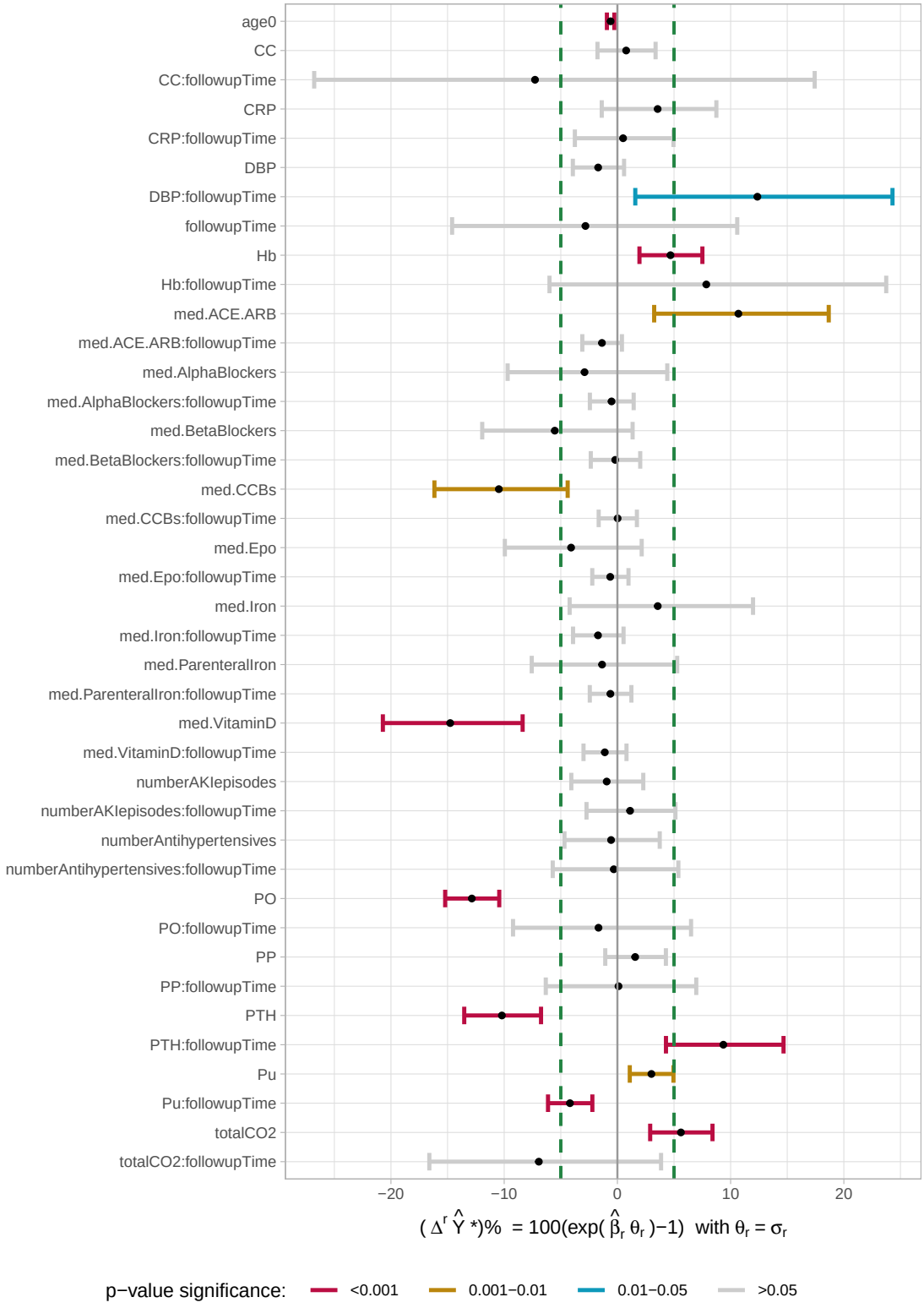

Figure 73: Relative change in eGFR for un-standardised model using 95% CIs: glomerulonephritis

A.7.4 Hypertensive kidney disease

Table 51: Estimated changes in outcome for changes in parameters for disease HKD

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\frac{\mathbb{E}(\Delta^r \hat{Y}^*)}{\mathbb{E}(\hat{Y}^*_{ij}) = 30}$
biochemical	CC	-0.1470	8.3e-02	0.1	-2.14	-0.64
	CC:followupTime	0.0113	2.3e-02	5.6	6.51	1.95
	Hb	0.0031	7.5e-04	17.1	5.40	1.62
	Hb:followupTime	0.0001	2.7e-04	304.5	4.41	1.32
	PO	-0.3039	5.2e-02	0.3	-8.27	-2.48
	PO:followupTime	-0.0206	1.5e-02	2.9	-5.77	-1.73
	PTH	-0.0057	1.6e-03	11.8	-6.53	-1.96
	PTH:followupTime	-0.0003	4.6e-04	52.3	-1.60	-0.48
	Pu	-0.0345	1.8e-02	1.6	-5.22	-1.57
	Pu:followupTime	-0.0163	4.6e-03	5.3	-8.24	-2.47
	totalCO2	0.0104	3.7e-03	3.5	3.73	1.12
	totalCO2:followupTime	-0.0009	1.1e-03	56.0	-4.91	-1.47
catagorical	comorbidityCancercurrent	0.0606	1.0e-01	1.0	6.24	1.87
	comorbidityCancercurrent:followupTime	0.0174	2.8e-02	1.0	1.75	0.53
	comorbidityCancerprevious	-0.1753	7.0e-02	1.0	-16.08	-4.82
	comorbidityCancerprevious:followupTime	0.0284	1.5e-02	1.0	2.88	0.86
	comorbidityCV1	-0.0991	3.8e-02	1.0	-9.43	-2.83
	comorbidityCV1:followupTime	0.0174	1.1e-02	1.0	1.75	0.53
	comorbidityCVover 1	-0.0810	4.1e-02	1.0	-7.78	-2.33
	comorbidityCVover 1:followupTime	0.0012	1.1e-02	1.0	0.12	0.04
	comorbidityOther	0.0377	4.7e-02	1.0	3.84	1.15
	comorbidityOther:followupTime	-0.0046	9.9e-03	1.0	-0.46	-0.14
	ethnicitynonWhite	0.1617	1.2e-01	1.0	17.55	5.26
	med.ACE.ARB	0.0138	3.1e-02	1.0	1.39	0.42
	med.ACE.ARB:followupTime	-0.0025	9.0e-03	1.0	-0.25	-0.07
	med.CCBs	-0.0413	3.1e-02	1.0	-4.04	-1.21
	med.CCBs:followupTime	0.0087	8.6e-03	1.0	0.88	0.26
	med.Epo	-0.0075	3.6e-02	1.0	-0.74	-0.22
	med.Epo:followupTime	-0.0052	1.1e-02	1.0	-0.52	-0.16
	med.Iron	0.0509	3.3e-02	1.0	5.22	1.57
	med.Iron:followupTime	-0.0139	1.2e-02	1.0	-1.38	-0.42
	med.ParenteralIron	-0.0261	3.1e-02	1.0	-2.58	-0.77
	med.ParenteralIron:followupTime	0.0106	1.0e-02	1.0	1.07	0.32
	med.VitaminD	-0.1115	3.3e-02	1.0	-10.55	-3.17
	med.VitaminD:followupTime	0.0102	1.1e-02	1.0	1.03	0.31
	smokingStatus0active	-0.1210	9.0e-02	1.0	-11.40	-3.42
	smokingStatus0ex-smoker	-0.1001	5.4e-02	1.0	-9.52	-2.86
	weeklyAlcohol01 to 14	-0.1232	5.3e-02	1.0	-11.59	-3.48
	weeklyAlcohol0over 14	-0.0767	6.8e-02	1.0	-7.38	-2.21
	age0	-0.0100	2.4e-03	1.0	-1.00	-0.30
	DBP	0.0013	9.2e-04	11.3	1.48	0.44
	DBP:followupTime	0.0001	3.1e-04	178.3	1.50	0.45
	followupTime	-0.0391	7.6e-02	1.0	-3.83	-1.15
	numberAKIepisodes	-0.0409	3.6e-02	0.4	-1.43	-0.43
	numberAKIepisodes:followupTime	-0.0011	1.2e-02	1.4	-0.16	-0.05
	numberAntihypertensives	-0.0391	1.4e-02	1.4	-5.36	-1.61
	numberAntihypertensives:followupTime	0.0076	4.0e-03	7.5	5.86	1.76

Table 51: Estimated changes in outcome for changes in parameters for disease HKD *(continued)*

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\hat{Y}^*_{ij}) = 30$
	numberClinicVisits	-0.0020	4.5e-03	2.4	-0.48	-0.14
	numberClinicVisits:followupTime	-0.0064	1.7e-03	10.4	-6.42	-1.93
	PP	0.0002	6.0e-04	19.3	0.39	0.12
	PP:followupTime	0.0002	1.9e-04	171.0	3.14	0.94

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

Figure 74: Relative change in eGFR for un-standardised model using 95% CIs: HKD

A.7.5 Other

Table 52: Estimated changes in outcome for changes in parameters for disease other

						$\mathbb{E}(\Delta^r \hat{Y}^*)$
category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\hat{Y}^*_{ij}) = 30$
biochemical	CC	-0.1869	8.6e-02	0.1	-2.72	-0.82
	CC:followupTime	0.0396	2.5e-02	5.6	24.74	7.42
	CRP	0.0009	6.6e-04	19.9	1.84	0.55
	CRP:followupTime	-0.0003	2.3e-04	71.0	-1.97	-0.59
	Hb	0.0057	8.6e-04	17.1	10.30	3.09
	Hb:followupTime	-0.0003	2.5e-04	304.5	-9.04	-2.71
	PO	-0.3167	5.5e-02	0.3	-8.60	-2.58
	PO:followupTime	-0.0448	1.4e-02	2.9	-12.14	-3.64
	PTH	-0.0087	1.4e-03	11.8	-9.85	-2.95
	PTH:followupTime	0.0012	3.5e-04	52.3	6.48	1.94
	totalCholesterol	0.0237	1.3e-02	1.1	2.76	0.83
	totalCholesterol:followupTime	-0.0019	3.7e-03	11.0	-2.04	-0.61
	totalCO2	0.0124	3.5e-03	3.5	4.50	1.35
	totalCO2:followupTime	0.0006	1.0e-03	56.0	3.32	1.00
catagorical	comorbidityCV1	-0.0227	3.7e-02	1.0	-2.24	-0.67
	comorbidityCV1:followupTime	-0.0167	9.7e-03	1.0	-1.65	-0.50
	comorbidityCVover 1	-0.0195	4.7e-02	1.0	-1.93	-0.58
	comorbidityCVover 1:followupTime	0.0098	1.3e-02	1.0	0.99	0.30
	comorbidityDiabetestype1	-0.1038	2.4e-01	1.0	-9.86	-2.96
	comorbidityDiabetestype1:followupTime	-0.0342	6.5e-02	1.0	-3.36	-1.01
	comorbidityDiabetestype2	0.0844	4.7e-02	1.0	8.81	2.64
	comorbidityDiabetestype2:followupTime	0.0033	1.1e-02	1.0	0.34	0.10
	med.BetaBlockers	-0.0468	3.7e-02	1.0	-4.57	-1.37
	med.BetaBlockers:followupTime	-0.0071	9.1e-03	1.0	-0.70	-0.21
	med.Epo	-0.0422	3.7e-02	1.0	-4.13	-1.24
	med.Epo:followupTime	-0.0186	1.1e-02	1.0	-1.84	-0.55
	med.Iron	-0.0477	4.0e-02	1.0	-4.66	-1.40
	med.Iron:followupTime	-0.0020	1.2e-02	1.0	-0.20	-0.06
	med.ParenteralIron	-0.0075	3.8e-02	1.0	-0.75	-0.22
	med.ParenteralIron:followupTime	-0.0169	1.1e-02	1.0	-1.68	-0.50
	med.VitaminD	-0.1557	3.9e-02	1.0	-14.42	-4.33
	med.VitaminD:followupTime	0.0115	1.1e-02	1.0	1.15	0.35
general	age0	-0.0077	1.7e-03	1.0	-0.76	-0.23
	followupTime	-0.0273	7.4e-02	1.0	-2.70	-0.81
	numberAKIepisodes	-0.0095	4.1e-02	0.4	-0.33	-0.10
	numberAKIepisodes:followupTime	-0.0090	1.2e-02	1.4	-1.28	-0.39
	numberClinicVisits	0.0102	4.2e-03	2.4	2.48	0.74
	numberClinicVisits:followupTime	-0.0010	1.2e-03	10.4	-1.01	-0.30

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

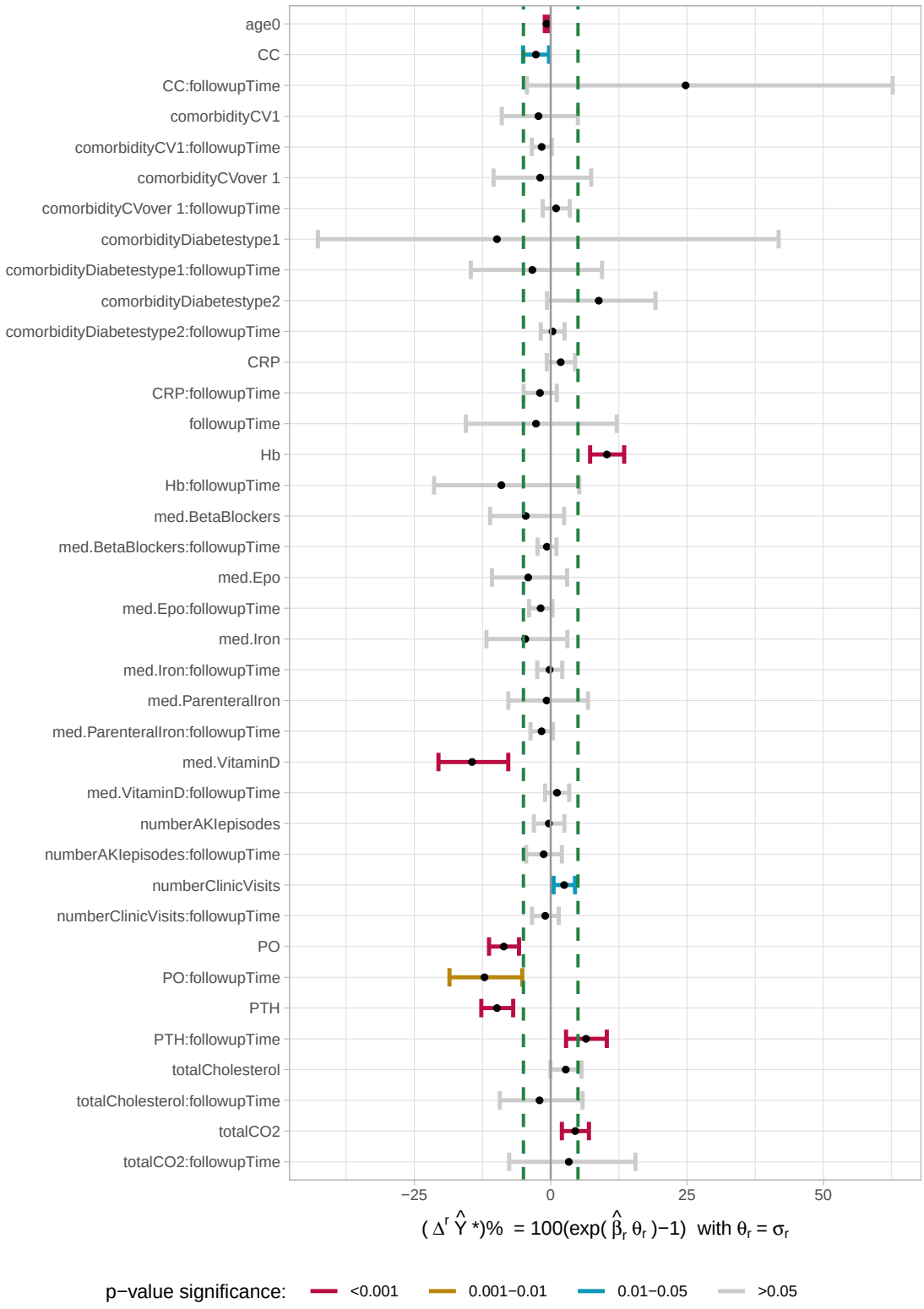

Figure 75: Relative change in eGFR for un-standardised model using 95% CIs: other

A.7.6 PKD

Table 53: Estimated changes in outcome for changes in parameters for disease PKD

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\Delta^r \hat{Y}^*)$ $\mathbb{E}(\hat{Y}^*_{ij}) = 30$
biochemical	CC	-0.1517	1.1e-01	0.1	-2.21	-0.66
	CC:followupTime	-0.0375	3.1e-02	5.6	-18.89	-5.67
	CRP	-0.0027	1.7e-03	19.9	-5.32	-1.60
	CRP:followupTime	0.0011	6.4e-04	71.0	8.46	2.54
	Hb	0.0023	1.3e-03	17.1	3.92	1.18
	Hb:followupTime	0.0009	3.7e-04	304.5	30.68	9.21
	PO	-0.1898	8.1e-02	0.3	-5.25	-1.57
	PO:followupTime	-0.0236	2.3e-02	2.9	-6.59	-1.98
	PTH	-0.0071	2.5e-03	11.8	-8.03	-2.41
	PTH:followupTime	0.0001	6.2e-04	52.3	0.34	0.10
	Pu	-0.0998	4.4e-02	1.6	-14.37	-4.31
	Pu:followupTime	0.0093	1.1e-02	5.3	5.02	1.51
	totalCO2	0.0135	5.2e-03	3.5	4.88	1.47
	totalCO2:followupTime	-0.0024	1.6e-03	56.0	-12.59	-3.78
catagorical	comorbidityCV1	0.0029	4.9e-02	1.0	0.29	0.09
	comorbidityCV1:followupTime	-0.0153	1.3e-02	1.0	-1.52	-0.46
	comorbidityCVover 1	-0.0832	9.1e-02	1.0	-7.99	-2.40
	comorbidityCVover 1:followupTime	0.0073	2.0e-02	1.0	0.74	0.22
	familyHistoryIHD0	-0.0645	8.8e-02	1.0	-6.24	-1.87
	med.ACE.ARB	0.0087	4.9e-02	1.0	0.87	0.26
	med.ACE.ARB:followupTime	-0.0069	1.2e-02	1.0	-0.69	-0.21
	med.CCBs	0.0856	4.7e-02	1.0	8.93	2.68
	med.CCBs:followupTime	-0.0147	1.2e-02	1.0	-1.46	-0.44
	med.Diuretics	0.0149	4.1e-02	1.0	1.50	0.45
	med.Diuretics:followupTime	-0.0188	1.3e-02	1.0	-1.86	-0.56
	med.Epo	0.0060	7.2e-02	1.0	0.60	0.18
	med.Epo:followupTime	-0.0260	1.8e-02	1.0	-2.57	-0.77
	med.Iron	-0.0658	7.0e-02	1.0	-6.37	-1.91
	med.Iron:followupTime	0.0170	2.0e-02	1.0	1.71	0.51
	med.ParenteralIron	-0.1571	6.4e-02	1.0	-14.54	-4.36
	med.ParenteralIron:followupTime	0.0245	1.8e-02	1.0	2.48	0.74
	med.VitaminD	-0.0877	5.7e-02	1.0	-8.40	-2.52
	med.VitaminD:followupTime	0.0109	1.2e-02	1.0	1.10	0.33
	occupation0ManagerialProfessional	0.1336	9.8e-02	1.0	14.29	4.29
	occupation0Intermediate	0.0491	1.2e-01	1.0	5.04	1.51
	occupation0NeverWorkedUnemployed	0.0082	2.0e-01	1.0	0.83	0.25
general	age0	-0.0190	3.5e-03	1.0	-1.88	-0.56
	DBP	0.0020	1.4e-03	11.3	2.30	0.69
	DBP:followupTime	-0.0001	4.2e-04	178.3	-1.77	-0.53
	followupTime	-0.0662	1.1e-01	1.0	-6.40	-1.92
	numberAntihypertensives	-0.0305	2.2e-02	1.4	-4.21	-1.26
	numberAntihypertensives:followupTime	0.0085	5.8e-03	7.5	6.61	1.98
	numberClinicVisits	0.0005	7.5e-03	2.4	0.11	0.03
	numberClinicVisits:followupTime	0.0000	1.8e-03	10.4	0.01	0.00
	PP	0.0024	1.1e-03	19.3	4.73	1.42
	PP:followupTime	0.0001	2.9e-04	171.0	1.55	0.47

Table 53: Estimated changes in outcome for changes in parameters for disease PKD *(continued)*

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\hat{Y}^*_{ij}) = 30$

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

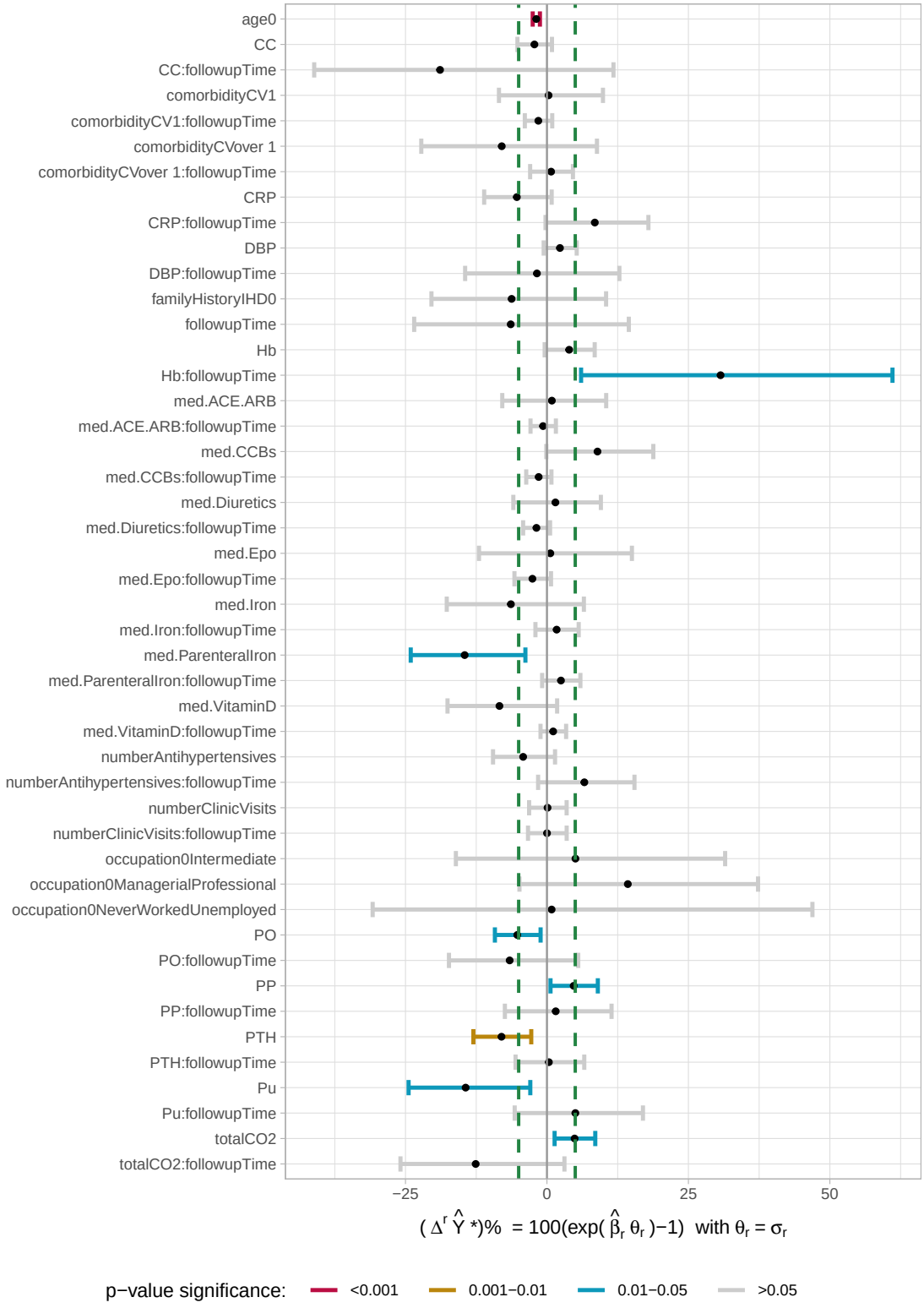

Figure 76: Relative change in eGFR for un-standardised model using 95% CIs: PKD

A.7.7 Pyelonephritis

Table 54: Estimated changes in outcome for changes in parameters for disease pyelonephritis

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\frac{\mathbb{E}(\Delta^r \hat{Y}^*)}{\mathbb{E}(\hat{Y}^*_{ij}) = 30}$
biochemical	CC	-0.1468	1.1e-01	0.1	-2.14	-0.64
	CC:followupTime	0.0052	3.1e-02	5.6	2.93	0.88
	Hb	0.0023	1.0e-03	17.1	3.98	1.19
	Hb:followupTime	0.0009	3.4e-04	304.5	30.72	9.22
	PO	-0.2828	6.7e-02	0.3	-7.72	-2.31
	PO:followupTime	-0.0204	1.9e-02	2.9	-5.74	-1.72
	PTH	-0.0019	2.3e-03	11.8	-2.25	-0.68
	PTH:followupTime	0.0005	4.2e-04	52.3	2.60	0.78
	Pu	0.0061	1.7e-02	1.6	0.95	0.29
	Pu:followupTime	-0.0218	5.3e-03	5.3	-10.91	-3.27
	totalCO2	0.0154	4.8e-03	3.5	5.61	1.68
	totalCO2:followupTime	-0.0031	1.3e-03	56.0	-15.73	-4.72
catagorical	comorbidityCancercurrent	-0.2215	1.2e-01	1.0	-19.87	-5.96
	comorbidityCancercurrent:followupTime	0.0481	2.9e-02	1.0	4.93	1.48
	comorbidityCancerprevious	-0.0708	1.1e-01	1.0	-6.83	-2.05
	comorbidityCancerprevious:followupTime	0.0443	2.3e-02	1.0	4.53	1.36
	comorbidityCV1	-0.0553	5.3e-02	1.0	-5.38	-1.61
	comorbidityCV1:followupTime	0.0076	1.3e-02	1.0	0.76	0.23
	comorbidityCVover 1	0.0026	7.2e-02	1.0	0.26	0.08
	comorbidityCVover 1:followupTime	-0.0194	1.5e-02	1.0	-1.92	-0.58
	comorbidityDiabetestype2	0.0616	7.5e-02	1.0	6.35	1.91
	comorbidityDiabetestype2:followupTime	0.0221	1.7e-02	1.0	2.24	0.67
	comorbidityGastrointestinal	0.0851	1.5e-01	1.0	8.88	2.66
	comorbidityGastrointestinal:followupTime	0.0227	2.1e-02	1.0	2.30	0.69
	familyHistoryIHD0	0.0570	7.7e-02	1.0	5.87	1.76
	med.ACE.ARB	0.0083	4.4e-02	1.0	0.84	0.25
	med.ACE.ARB:followupTime	-0.0091	1.2e-02	1.0	-0.91	-0.27
	med.AlphaBlockers	-0.0607	5.9e-02	1.0	-5.89	-1.77
	med.AlphaBlockers:followupTime	0.0044	1.5e-02	1.0	0.44	0.13
	med.CCBs	0.0521	5.0e-02	1.0	5.35	1.61
	med.CCBs:followupTime	-0.0079	1.2e-02	1.0	-0.79	-0.24
	med.Epo	-0.0960	6.4e-02	1.0	-9.15	-2.75
	med.Epo:followupTime	0.0142	1.9e-02	1.0	1.43	0.43
	med.Iron	-0.0292	6.9e-02	1.0	-2.87	-0.86
	med.Iron:followupTime	0.0026	1.5e-02	1.0	0.26	0.08
	med.ParenteralIron	0.0051	4.6e-02	1.0	0.51	0.15
	med.ParenteralIron:followupTime	-0.0329	2.2e-02	1.0	-3.23	-0.97
	med.VitaminD	-0.1185	4.3e-02	1.0	-11.17	-3.35
	med.VitaminD:followupTime	0.0159	1.0e-02	1.0	1.60	0.48
	occupation0ManagerialProfessional	-0.1897	1.0e-01	1.0	-17.28	-5.18
	occupation0Intermediate	0.1509	9.9e-02	1.0	16.28	4.88
	occupation0NeverWorkedUnemployed	-0.0892	1.7e-01	1.0	-8.53	-2.56
	sexfemale	-0.1501	8.0e-02	1.0	-13.94	-4.18
	smokingStatus0active	-0.1343	1.2e-01	1.0	-12.56	-3.77
	smokingStatus0ex-smoker	0.0052	8.8e-02	1.0	0.52	0.16
	weeklyAlcohol01 to 14	-0.1031	8.9e-02	1.0	-9.79	-2.94
	weeklyAlcohol0over 14	0.2773	1.2e-01	1.0	31.96	9.59

Table 54: Estimated changes in outcome for changes in parameters for disease pyelonephritis *(continued)*

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\hat{Y}^*_{ij}) = 30$
general	age0	-0.0081	2.4e-03	1.0	-0.81	-0.24
	bodyMassIndex	0.0037	5.3e-03	5.9	2.20	0.66
	bodyMassIndex:followupTime	0.0013	9.6e-04	69.8	9.74	2.92
	DBP	0.0020	1.2e-03	11.3	2.31	0.69
	DBP:followupTime	-0.0003	3.7e-04	178.3	-5.57	-1.67
	followupTime	-0.0892	9.4e-02	1.0	-8.53	-2.56
	numberAKIepisodes	-0.0926	6.8e-02	0.4	-3.21	-0.96
	numberAKIepisodes:followupTime	0.0072	2.4e-02	1.4	1.04	0.31
	numberAntihypertensives	-0.0748	2.5e-02	1.4	-10.01	-3.00
	numberAntihypertensives:followupTime	0.0027	6.5e-03	7.5	2.07	0.62
	numberClinicVisits	0.0021	7.0e-03	2.4	0.50	0.15
	numberClinicVisits:followupTime	0.0003	2.5e-03	10.4	0.34	0.10
	PP	-0.0001	1.0e-03	19.3	-0.26	-0.08
	PP:followupTime	0.0004	2.9e-04	171.0	6.88	2.06

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

Figure 77: Relative change in eGFR for un-standardised model using 95% CIs: pyelonephritis

A.7.8 Renovascular

Table 55: Estimated changes in outcome for changes in parameters for disease renovascular

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\Delta^r \hat{Y}^*)$ / $\mathbb{E}(\hat{Y}^*_{ij}) = 30$
biochemical	CC	-0.1076	1.4e-01	0.1	-1.57	-0.47
	CC:followupTime	0.0025	3.8e-02	5.6	1.41	0.42
	CRP	0.0013	8.4e-04	19.9	2.59	0.78
	CRP:followupTime	-0.0003	2.5e-04	71.0	-2.18	-0.65
	Hb	0.0031	1.1e-03	17.1	5.40	1.62
	Hb:followupTime	0.0004	4.0e-04	304.5	11.38	3.41
	PO	-0.4033	6.8e-02	0.3	-10.82	-3.25
	PO:followupTime	0.0213	2.2e-02	2.9	6.35	1.91
	PTH	-0.0071	2.7e-03	11.8	-8.06	-2.42
	PTH:followupTime	0.0010	9.9e-04	52.3	5.47	1.64
	Pu	-0.0169	3.4e-02	1.6	-2.60	-0.78
	Pu:followupTime	-0.0079	6.4e-03	5.3	-4.08	-1.22
	totalCholesterol	0.0072	1.6e-02	1.1	0.83	0.25
	totalCholesterol:followupTime	-0.0143	5.3e-03	11.0	-14.59	-4.38
catagorical	comorbidityCV1	-0.0688	7.2e-02	1.0	-6.64	-1.99
	comorbidityCV1:followupTime	0.0262	2.0e-02	1.0	2.66	0.80
	comorbidityCVover 1	-0.1726	7.0e-02	1.0	-15.86	-4.76
	comorbidityCVover 1:followupTime	0.0122	1.9e-02	1.0	1.22	0.37
	comorbidityDiabetestype1	-0.0527	2.1e-01	1.0	-5.13	-1.54
	comorbidityDiabetestype1:followupTime	0.0317	5.2e-02	1.0	3.22	0.97
	comorbidityDiabetestype2	0.0450	5.0e-02	1.0	4.60	1.38
	comorbidityDiabetestype2:followupTime	0.0080	1.3e-02	1.0	0.81	0.24
	comorbidityGastrointestinal	0.1229	8.6e-02	1.0	13.08	3.92
	comorbidityGastrointestinal:followupTime	-0.0217	2.0e-02	1.0	-2.15	-0.65
	familyHistoryIHD0	0.0449	6.5e-02	1.0	4.59	1.38
	med.ACE.ARB	-0.0119	4.0e-02	1.0	-1.19	-0.36
	med.ACE.ARB:followupTime	0.0164	1.2e-02	1.0	1.66	0.50
	med.AlphaBlockers	0.0054	3.9e-02	1.0	0.54	0.16
	med.AlphaBlockers:followupTime	0.0341	1.3e-02	1.0	3.47	1.04
	med.Diuretics	-0.0889	3.6e-02	1.0	-8.50	-2.55
	med.Diuretics:followupTime	0.0065	1.1e-02	1.0	0.66	0.20
	med.Epo	-0.1739	5.0e-02	1.0	-15.96	-4.79
	med.Epo:followupTime	-0.0124	1.7e-02	1.0	-1.23	-0.37
	med.Iron	-0.0300	3.9e-02	1.0	-2.95	-0.89
	med.Iron:followupTime	-0.0183	1.1e-02	1.0	-1.81	-0.54
	med.ParenteralIron	0.0935	4.7e-02	1.0	9.80	2.94
	med.ParenteralIron:followupTime	-0.0328	1.7e-02	1.0	-3.22	-0.97
	med.VitaminD	-0.0931	4.2e-02	1.0	-8.89	-2.67
	med.VitaminD:followupTime	-0.0153	1.6e-02	1.0	-1.52	-0.46
	occupation0ManagerialProfessional	0.0590	8.3e-02	1.0	6.08	1.82
	occupation0Intermediate	0.0224	9.1e-02	1.0	2.27	0.68
	occupation0NeverWorkedUnemployed	0.3974	4.0e-01	1.0	48.80	14.64
	smokingStatus0active	-0.2123	1.2e-01	1.0	-19.12	-5.74
	smokingStatus0ex-smoker	-0.1537	8.9e-02	1.0	-14.25	-4.28
	weeklyAlcohol01 to 14	-0.0555	8.0e-02	1.0	-5.40	-1.62
	weeklyAlcohol0over 14	0.1141	9.5e-02	1.0	12.09	3.63
	age0	-0.0109	4.2e-03	1.0	-1.08	-0.32

Table 55: Estimated changes in outcome for changes in parameters for disease renovascular *(continued)*

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\hat{Y}_{ij}^*) = 30$
general	DBP	0.0005	1.2e-03	11.3	0.60	0.18
	DBP:followupTime	0.0009	3.7e-04	178.3	18.40	5.52
	followupTime	-0.0970	1.1e-01	1.0	-9.24	-2.77
	numberAntihypertensives	0.0037	1.7e-02	1.4	0.52	0.16
	numberAntihypertensives:followupTime	-0.0064	5.3e-03	7.5	-4.67	-1.40
	PP	0.0019	6.7e-04	19.3	3.70	1.11
	PP:followupTime	-0.0001	2.1e-04	171.0	-2.50	-0.75

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}_{ij}^*)$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

Figure 78: Relative change in eGFR for un-standardised model using 95% CIs: renovascular

A.7.9 Unknown disease

Table 56: Estimated changes in outcome for changes in parameters for disease unknown

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\frac{\mathbb{E}(\Delta^r \hat{Y}^*)}{\mathbb{E}(\hat{Y}^*_{ij}) = 30}$
biochemical	CC	-0.1060	9.8e-02	0.1	-1.55	-0.47
	CC:followupTime	-0.0237	2.9e-02	5.6	-12.39	-3.72
	Hb	0.0036	1.0e-03	17.1	6.42	1.93
	Hb:followupTime	0.0001	3.3e-04	304.5	3.67	1.10
	PO	-0.3078	6.2e-02	0.3	-8.37	-2.51
	PO:followupTime	-0.0787	2.0e-02	2.9	-20.34	-6.10
	PTH	-0.0123	2.0e-03	11.8	-13.51	-4.05
	PTH:followupTime	0.0020	5.7e-04	52.3	11.27	3.38
	Pu	-0.0342	1.4e-02	1.6	-5.18	-1.55
	Pu:followupTime	-0.0018	5.9e-03	5.3	-0.93	-0.28
	totalCholesterol	0.0015	1.5e-02	1.1	0.17	0.05
	totalCholesterol:followupTime	0.0091	5.1e-03	11.0	10.48	3.14
	totalCO2	0.0154	4.5e-03	3.5	5.59	1.68
	totalCO2:followupTime	-0.0023	1.5e-03	56.0	-12.17	-3.65
catagorical	comorbidityCancercurrent	0.1649	1.3e-01	1.0	17.93	5.38
	comorbidityCancercurrent:followupTime	0.0123	5.2e-02	1.0	1.23	0.37
	comorbidityCancerprevious	0.0629	6.8e-02	1.0	6.49	1.95
	comorbidityCancerprevious:followupTime	-0.0224	1.9e-02	1.0	-2.21	-0.66
	med.ACE.ARB	-0.0189	3.8e-02	1.0	-1.88	-0.56
	med.ACE.ARB:followupTime	-0.0217	1.3e-02	1.0	-2.14	-0.64
	med.BetaBlockers	0.0093	4.1e-02	1.0	0.93	0.28
	med.BetaBlockers:followupTime	-0.0163	1.3e-02	1.0	-1.62	-0.49
	med.CCBs	0.0138	3.6e-02	1.0	1.39	0.42
	med.CCBs:followupTime	-0.0247	1.3e-02	1.0	-2.44	-0.73
	med.Epo	-0.0755	3.8e-02	1.0	-7.27	-2.18
	med.Epo:followupTime	0.0075	1.4e-02	1.0	0.76	0.23
	med.Iron	-0.0240	4.0e-02	1.0	-2.37	-0.71
	med.Iron:followupTime	0.0151	1.3e-02	1.0	1.52	0.45
	med.ParenteralIron	-0.0176	3.9e-02	1.0	-1.75	-0.52
	med.ParenteralIron:followupTime	0.0110	1.7e-02	1.0	1.11	0.33
	med.VitaminD	-0.1207	4.3e-02	1.0	-11.37	-3.41
	med.VitaminD:followupTime	-0.0134	1.3e-02	1.0	-1.33	-0.40
	weeklyAlcohol01 to 14	-0.0494	5.2e-02	1.0	-4.82	-1.45
	weeklyAlcohol0over 14	0.0834	6.4e-02	1.0	8.69	2.61
general	age0	-0.0067	1.8e-03	1.0	-0.67	-0.20
	DBP	0.0024	1.2e-03	11.3	2.71	0.81
	DBP:followupTime	-0.0005	4.3e-04	178.3	-8.92	-2.67
	followupTime	0.1521	9.4e-02	1.0	16.43	4.93
	numberAKIepisodes	0.1381	5.8e-02	0.4	4.99	1.50
	numberAKIepisodes:followupTime	-0.0190	1.7e-02	1.4	-2.70	-0.81
	numberAntihypertensives	-0.0335	1.8e-02	1.4	-4.62	-1.39
	numberAntihypertensives:followupTime	0.0147	6.1e-03	7.5	11.60	3.48
	PP	0.0000	7.2e-04	19.3	0.02	0.01
	PP:followupTime	0.0000	2.5e-04	171.0	-0.71	-0.21

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

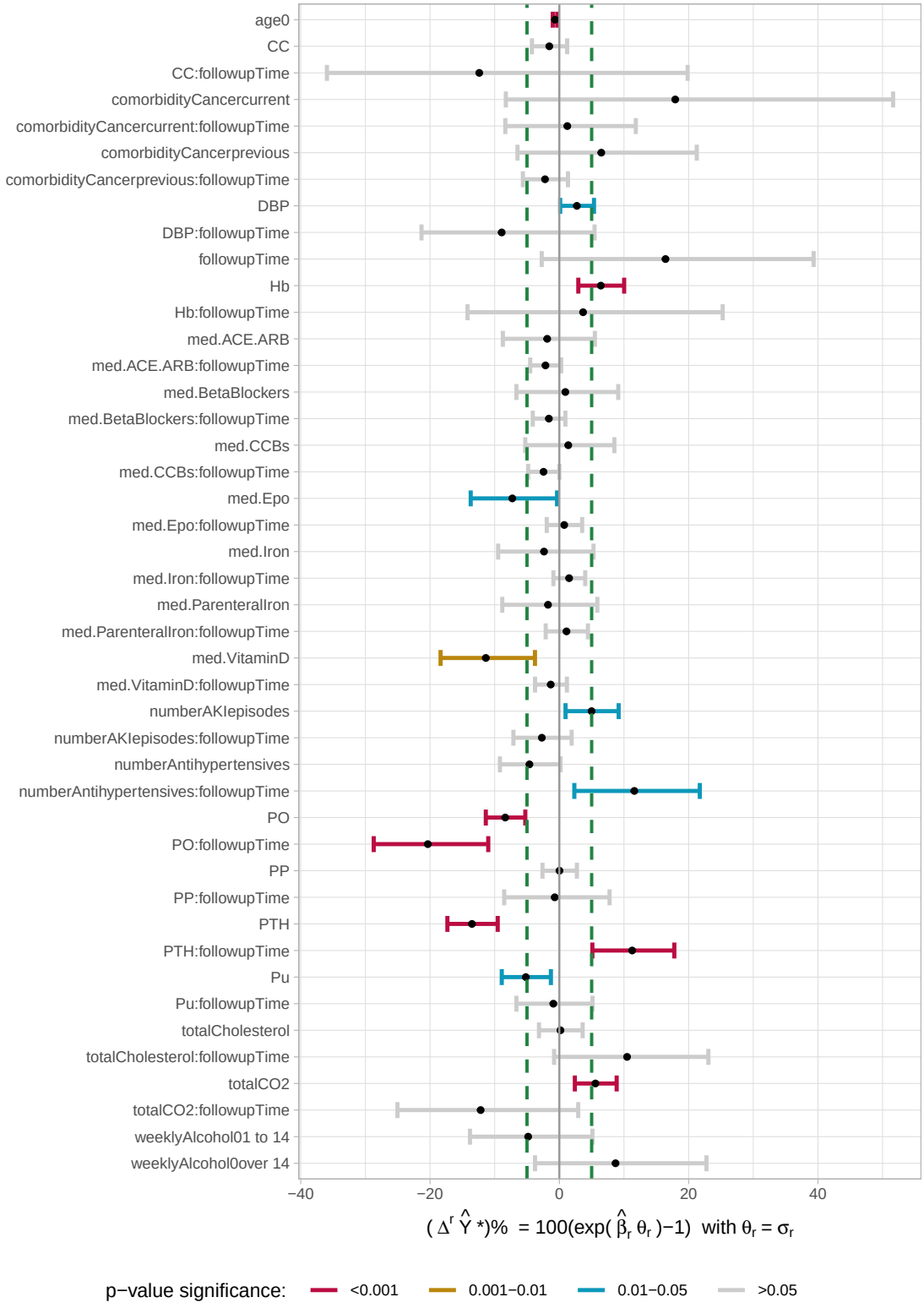

Figure 79: Relative change in eGFR for un-standardised model using 95% CIs: unknown

A.7.10 Single model all diseases

Table 57: Estimated changes in outcome for changes in parameters for single model all diseases

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\frac{\mathbb{E}(\Delta^r \hat{Y}^*)}{\mathbb{E}(\hat{Y}^*_{ij}) = 30}$
biochemical	CC	-0.1071	3.4e-02	0.1	-1.57	-0.47
	CC:followupTime	-0.0050	9.3e-03	5.6	-2.73	-0.82
	Hb	0.0033	3.3e-04	17.1	5.74	1.72
	Hb:followupTime	0.0003	1.0e-04	304.5	10.89	3.27
	PO	-0.3825	2.1e-02	0.3	-10.29	-3.09
	PO:followupTime	-0.0186	5.9e-03	2.9	-5.23	-1.57
	PTH	-0.0074	5.6e-04	11.8	-8.34	-2.50
	PTH:followupTime	0.0008	1.3e-04	52.3	4.39	1.32
	Pu	0.0021	3.7e-03	1.6	0.33	0.10
	Pu:followupTime	-0.0075	1.1e-03	5.3	-3.88	-1.16
	totalCholesterol	0.0102	4.9e-03	1.1	1.18	0.35
	totalCholesterol:followupTime	-0.0012	1.5e-03	11.0	-1.35	-0.41
	totalCO2	0.0142	1.5e-03	3.5	5.17	1.55
	totalCO2:followupTime	-0.0017	4.3e-04	56.0	-9.27	-2.78
	comorbidityCancercurrent	0.0259	3.8e-02	1.0	2.62	0.79
	comorbidityCancercurrent:followupTime	-0.0009	9.8e-03	1.0	-0.09	-0.03
	comorbidityCancerprevious	-0.0488	2.5e-02	1.0	-4.76	-1.43
	comorbidityCancerprevious:followupTime	0.0056	6.2e-03	1.0	0.56	0.17
	comorbidityCV1	-0.0329	1.5e-02	1.0	-3.23	-0.97
	comorbidityCV1:followupTime	0.0003	4.3e-03	1.0	0.03	0.01
	comorbidityCVover 1	-0.0485	1.8e-02	1.0	-4.74	-1.42
	comorbidityCVover 1:followupTime	0.0059	4.7e-03	1.0	0.60	0.18
	disease diabetic nephropathy	-0.0963	3.2e-02	1.0	-9.18	-2.75
	disease glomerulonephritis	0.0091	3.4e-02	1.0	0.91	0.27
	disease HKD	-0.0536	3.5e-02	1.0	-5.22	-1.57
	disease obstruction	-0.3433	8.5e-02	1.0	-29.05	-8.72
	disease polycystic kidney disease	-0.1782	4.4e-02	1.0	-16.32	-4.90
	disease pyelonephritis	-0.1395	4.4e-02	1.0	-13.02	-3.91
	disease renovascular disease	-0.0197	4.1e-02	1.0	-1.95	-0.58
	disease unknown	-0.0556	3.5e-02	1.0	-5.41	-1.62
	med.ACE.ARB	0.0447	1.4e-02	1.0	4.57	1.37
	med.ACE.ARB:followupTime	-0.0063	3.8e-03	1.0	-0.63	-0.19
	med.AlphaBlockers	-0.0169	1.4e-02	1.0	-1.67	-0.50
	med.AlphaBlockers:followupTime	-0.0052	4.2e-03	1.0	-0.52	-0.16
	med.CCBs	-0.0311	1.3e-02	1.0	-3.06	-0.92
	med.CCBs:followupTime	-0.0007	3.8e-03	1.0	-0.07	-0.02
	med.Diuretics	-0.0426	1.3e-02	1.0	-4.17	-1.25
	med.Diuretics:followupTime	0.0058	3.9e-03	1.0	0.58	0.17
	med.Epo	-0.0745	1.3e-02	1.0	-7.18	-2.15
	med.Epo:followupTime	-0.0052	3.9e-03	1.0	-0.52	-0.16
	med.Iron	0.0077	1.4e-02	1.0	0.77	0.23
	med.Iron:followupTime	-0.0090	4.3e-03	1.0	-0.90	-0.27
	med.Other	-0.0507	2.4e-02	1.0	-4.95	-1.48
	med.Other:followupTime	0.0133	7.3e-03	1.0	1.34	0.40
	med.ParenteralIron	-0.0228	1.3e-02	1.0	-2.26	-0.68
	med.ParenteralIron:followupTime	0.0007	4.1e-03	1.0	0.07	0.02
	med.VitaminD	-0.1429	1.4e-02	1.0	-13.32	-4.00

Table 57: Estimated changes in outcome for changes in parameters for single model all diseases *(continued)*

category	parameter	$\hat{\beta}_r$	se	θ_r	$(\Delta^r \hat{Y}^*)\%$	$\mathbb{E}(\hat{Y}^*_{ij}) = 30$
	med.VitaminD:followupTime	0.0064	3.8e-03	1.0	0.64	0.19
general	age0	-0.0070	7.1e-04	1.0	-0.69	-0.21
	DBP	0.0010	3.9e-04	11.3	1.10	0.33
	DBP:followupTime	0.0001	1.2e-04	178.3	1.80	0.54
	followupTime	-0.0079	3.0e-02	1.0	-0.78	-0.24
	numberAKIepisodes	-0.0134	1.5e-02	0.4	-0.47	-0.14
	numberAKIepisodes:followupTime	-0.0030	3.7e-03	1.4	-0.44	-0.13
	numberAntihypertensives	-0.0088	6.9e-03	1.4	-1.24	-0.37
	numberAntihypertensives:followupTime	-0.0008	1.9e-03	7.5	-0.60	-0.18
	numberClinicVisits	-0.0025	1.7e-03	2.4	-0.60	-0.18
	numberClinicVisits:followupTime	-0.0002	4.7e-04	10.4	-0.26	-0.08
	PP	0.0006	2.5e-04	19.3	1.24	0.37
	PP:followupTime	0.0000	7.7e-05	171.0	0.21	0.06

Note: both $\mathbb{E}(\Delta^r \hat{Y}^*)$ and $\mathbb{E}(\hat{Y}^*_{ij})$ have units mL/min/1.73m^2 and $\theta_r = \sigma_r$

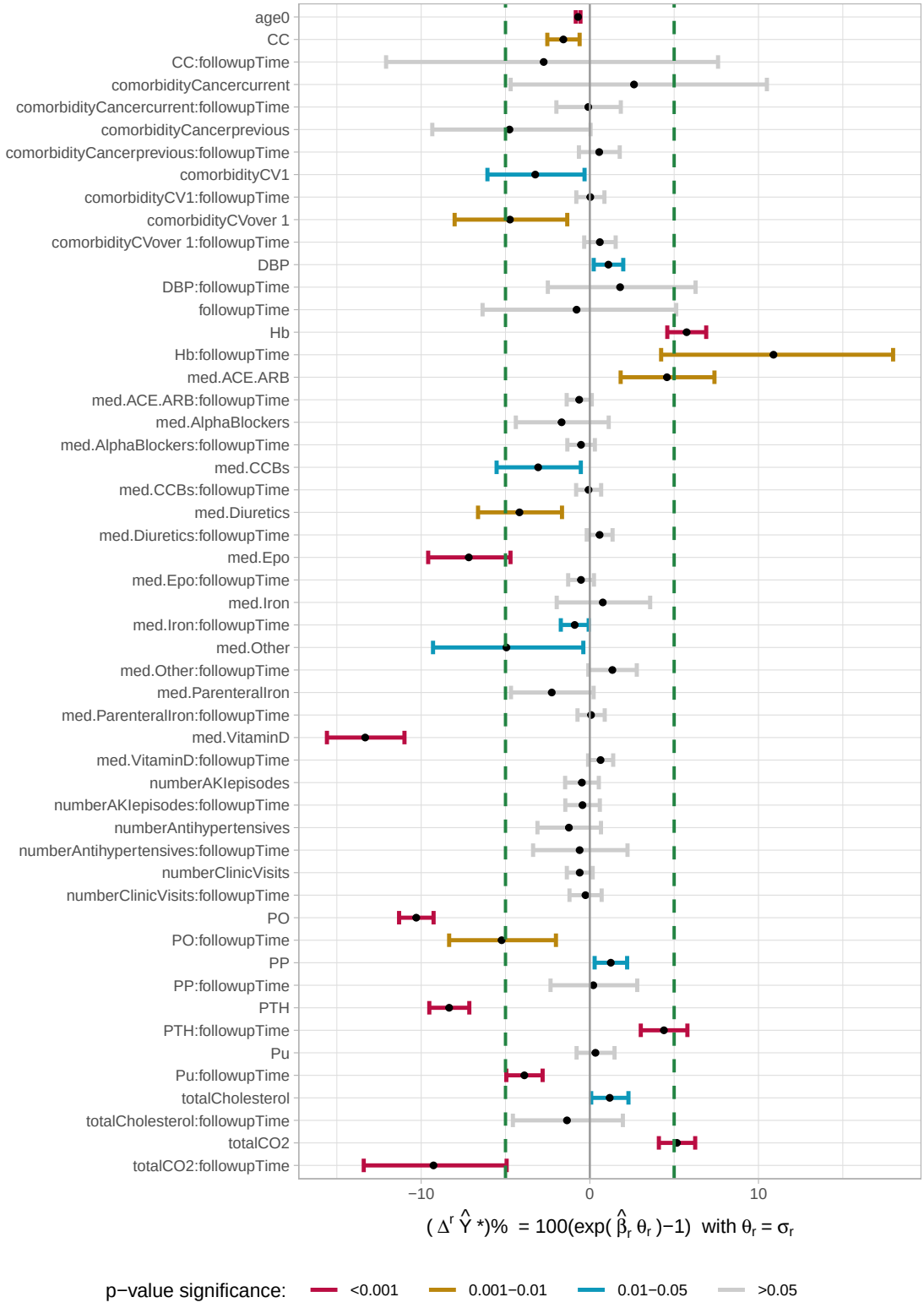

Figure 80: Relative change in eGFR for un-standardised model using 95% CIs for single model all diseases