University of Dundee

Performance of prediction models for nephropathy in people with type 2 diabetes

Slieker, Roderick C.; van der Heijden, Amber A.; Siddiqui, Moneeza K.; Langendoen-Gort, Marlous; Nijpels, Giel ; Herings, Ron

Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and external validation study

Roderick C Slieker,[1,2] Amber A W A van der Heijden,[3] Moneeza K Siddiqui,[4] Marlous Langendoen-Gort,[3] Giel Nijpels,[3] Ron Herings,[1,5] Talitha L Feenstra,[6,7] Karel G M Moons,[8,9] Samira Bell,[4] Petra J Elders,[3] Leen M 't Hart,[1,2,10] Joline W J Beulens[1,8]

ABSTRACT

OBJECTIVES
To identify and assess the quality and accuracy of prognostic models for nephropathy and to validate these models in external cohorts of people with type 2 diabetes.

DESIGN
Systematic review and external validation.

DATA SOURCES
PubMed and Embase.

ELIGIBILITY CRITERIA
Studies describing the development of a model to predict the risk of nephropathy, applicable to people with type 2 diabetes.

METHODS
Screening, data extraction, and risk of bias assessment were done in duplicate. Eligible models were externally validated in the Hoorn Diabetes Care System (DCS) cohort (n=11 450) for the same outcomes for which they were developed. Risks of nephropathy were calculated and compared with observed risk over 2, 5, and 10 years of follow-up. Model performance was assessed based on intercept adjusted calibration and discrimination (Harrell's C statistic).

RESULTS
41 studies included in the systematic review reported 64 models, 46 of which were developed in a population with diabetes and 18 in the general population including diabetes as a predictor. The predicted outcomes included albuminuria, diabetic kidney disease, chronic kidney disease (general population), and end stage renal disease. The reported apparent discrimination of the 46 models varied considerably across the different predicted outcomes, from 0.60 (95% confidence interval 0.56 to 0.64) to 0.99 (not available) for the models developed in a diabetes population and from 0.59 (not available) to 0.96 (0.95 to 0.97) for the models developed in the general population. Calibration was reported in 31 of the 41 studies, and the models were generally well calibrated. 21 of the 64 retrieved models were externally validated in the Hoorn DCS cohort for predicting risk of albuminuria, diabetic kidney disease, and chronic kidney disease, with considerable variation in performance across prediction horizons and models. For all three outcomes, however, at least two models had C statistics >0.8, indicating excellent discrimination. In a secondary external validation in GoDARTS (Genetics of Diabetes Audit and Research in Tayside Scotland), models developed for diabetic kidney disease outperformed those for chronic kidney disease. Models were generally well calibrated across all three prediction horizons.

CONCLUSIONS
This study identified multiple prediction models to predict albuminuria, diabetic kidney disease, chronic kidney disease, and end stage renal disease. In the external validation, discrimination and calibration for albuminuria, diabetic kidney disease, and chronic kidney disease varied considerably across prediction horizons and models. For each outcome, however, specific models showed good discrimination and calibration across the three prediction horizons, with clinically accessible predictors, making them applicable in a clinical setting.

SYSTEMATIC REVIEW REGISTRATION
PROSPERO CRD42020192831.

WHAT IS ALREADY KNOWN ON THIS TOPIC

People with type 2 diabetes are at increased risk of nephropathy

Many studies developed prediction models to identify those at high risk of nephropathy in people with type 2 diabetes or with diabetes as a predictor in the model

These models can be used in clinical practice provided that they are also accurate in external target populations

WHAT THIS STUDY ADDS

This study identified many studies that developed prognostic models to predict nephropathy in people with type 2 diabetes or with diabetes as a risk factor with variable performance across time horizons and models

Several models showed good performance in people with type 2 diabetes across various prediction horizons for different nephropathy outcomes,

Models, however, performed best at the two year horizon and for diabetic kidney disease as outcome

Introduction

People with type 2 diabetes are at high risk of microvascular and macrovascular complications.[1][2] About 25-35% of people with type 2 diabetes develop nephropathy,[3-5] one of the leading causes of end stage renal disease, which is associated with a low quality of life and high mortality.[6-9] As renal histological changes might already be advanced by the time a decline in renal function is detected,[10-12] early identification of those at risk is essential to initiate targeted preventive treatment. Moreover, interventions are more effective at earlier than more advanced stages

of nephropathy.[7] [13] Decisions for interventions can be achieved by regularly estimating an individual's risk of nephropathy based on validated prediction models. These prediction models should preferably be based on routine clinical markers and not expensive biomarkers, given that routine and low cost markers will be easier to implement into clinical practice.

Preventive treatment includes intensified glycaemic control, nephroprotective glucose lowering drugs such as glucagon-like peptide 1 receptor antagonists and sodium-glucose cotransporter 2 inhibitors, blood pressure control, and inhibition of the renin-angiotensin system.[14-17] Moreover, treatment with glucose lowering drugs such as sodium-glucose cotransporter 2 inhibitors has been suggested to prevent the progression of renal disease.[18] Treatment of end stage renal disease comprises renal replacement therapy with dialysis and transplantation.[7] [19] The efficacy of renal replacement therapy, however, has been shown to be worse in people with diabetes.[4]

Prediction models for nephropathy allow the accurate estimation of nephropathy risk in people with type 2 diabetes. To ensure reliability though, developed prediction models need to be externally validated in their targeted populations, and ideally compared head to head on their predictive performance.[20] [21] Only when proven valid in external validation, can prediction models be considered relevant to clinical practice to improve decision making and improve the cost effectiveness of care. Several prediction models for renal impairment and nephropathy in people with type 2 diabetes or applicable to this population have been developed.

Known risk factors for diabetic nephropathy include urinary albumin excretion, glucose levels, blood pressure, dyslipidaemia, obesity, smoking, duration of diabetes, age, sex, and retinopathy.[22] More recently, identified risk factors include oxidative stress, inflammation, genetic background, ethnicity, and glomerular hyperfiltration.[22]

Several systematic reviews on prognostic models for nephropathy have been performed,[23-25] although several years ago[23] [25] [26] or focusing only on the general population.[24] Also, although some external validation of the models has been done, it was either limited to the general population[24] [26] [27] or only included a small number of (other) models as part of a model development paper.[28] Given that people with type 2 diabetes have an increased risk of developing nephropathy, it is important to assess the performance of prognostic models in a head-to-head comparison in a large scale population of those with type 2 diabetes. We therefore systematically reviewed existing prediction models for nephropathy applicable to people with type 2 diabetes. The retrieved models were appraised on quality and subsequently validated and compared head to head in the Hoorn DCS cohort, a prospective cohort of more than 14 000 people from the Hoorn region in the Netherlands.[29] Finally, we validated the top performing models in GoDARTS (Genetics of Diabetes Audit and Research in Tayside Scotland), a secondary prospective cohort from Scotland.[30]

Methods

We carried out a systematic review and external validation study. This systematic review was performed according to guidance of the Cochrane Prognosis Methods Group (methods.cochrane.org/prognosis) and reported to the criteria of the preferred reporting items for systematic reviews and meta-analyses statement and a measurement tool to assess systematic reviews guidelines.[31-33] The external validation study was reported in line with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis guideline (supplementary tables S1 and S2).[34] [35]

Search strategy

We systematically searched PubMed and Embase from inception until 16 June 2020 to identify prediction models for the risk of developing nephropathy applicable to people with type 2 diabetes. This systematic literature search was based on a predefined search string that was specifically developed for both databases (supplementary table S3). Online software, covidence (Melbourne, Australia), was used to perform a systematic review of identified development studies.

Study selection

Four researchers (MLG, AAvdH, RCS, and JWJB) independently reviewed in duplicate the title, abstract, and full text article of retrieved studies. Disagreements between two reviewers were resolved by consensus with a third reviewer. A study was included when the prediction model was developed in a population in which the majority of people had a diagnosis of type 2 diabetes, or in the general population, and included diabetes as a predictor; the risk of developing nephropathy could be calculated from the reported prediction model or rule; nephropathy was the outcome of the model; and follow-up was longer than one year. We also included studies for which the type of diabetes was not specified but was probably type 2 diabetes based on the population age. Excluded studies were those performed in populations restricted to other forms of diabetes. A study was also excluded when it was conducted in animals, it was not written in English or Dutch, the prediction model was developed in a population with other physical severe conditions or in a postsurgical population, a new predictor was added to an original model, or the model only consisted of one predictor. In addition, reference lists of all screened full text articles and relevant systematic reviews were assessed for additional eligible studies. Any study type other than cross sectional (ie, without patient follow-up) was included—namely, randomised trials, cohort studies, and registry based studies.

Data extraction

Data from the included studies were systematically extracted according to the criteria of the checklist for

critical appraisal and data extraction for systematic reviews of prediction modelling studies checklist.[36] This checklist includes specifics about the research design, study population, outcome of the model, predictors, sample size, missing values, model development, performance of the model, model validation, results, and interpretation of the model. MLG, JWJB, and AAvdH extracted data on all these domains in duplicate from the included studies.

Quality assessment

Risk of bias and concern for applicability was assessed with the prediction model risk of bias assessment tool, which was specifically developed to assess diagnostic and prognostic prediction models.[37 38] This tool consists of 20 items and includes specifics about the research design, study population, outcome of the model, predictors, handling of the data, and performance measures. MLG, AAvdH, and JWJB independently extracted data and assessed the quality of each included study in duplicate. Any disagreements were resolved in consensus meetings.

External validation

External validation of the retrieved models was performed in the Hoorn Diabetes Care System cohort, a prospective cohort study from the Hoorn region in the Netherlands.[25] People with type 2 diabetes visit the Diabetes Care System in Hoorn once a year to have their diabetes monitored and multiple standardised measurements taken. Individuals were able to opt-out for the anonymous use of their data. We considered the year of entry to the study as baseline and used data until 31 December 2017 in the current study. The data of 12 155 people were included. Age at diagnosis was not imputed, and those with missing diagnosis dates were excluded (n=705), resulting in 11 450 individuals for analysis. Age at diabetes diagnosis was based on registry data. All laboratory measurements were measured after the participant had fasted. Glycated haemoglobin (HbA_{1c}) was assessed by turbidimetric inhibition immunoassay of haemolysed whole EDTA blood. Blood glucose level was assessed in fluorinated plasma with the ultraviolet test using hexokinase. Levels of triglycerides, total cholesterol, and high density lipoprotein cholesterol were determined enzymatically (Cobas c501; Roche Diagnostics, Mannheim, Germany). Low density lipoprotein cholesterol levels were calculated. Albumin was determined by the antigen's reaction with anti-albumin antibodies and measured turbidimetrically from an overnight first voided urine sample. Creatinine concentrations were determined enzymatically in heparinised plasma and urine (Cobas c501; Roche Diagnostics).

Presence of retinopathy was based on the EURODIAB 0-5 scale, with grade 0 representing no retinopathy; grade 1, minimal non-proliferative retinopathy; grade 2, moderate non-proliferative retinopathy; grade 3, severe non-proliferative or pre-proliferative retinopathy; grade 4, photocoagulated retinopathy;

and grade 5, proliferative retinopathy.[39] Blood pressure was measured twice at an interval of three minutes using a random zero sphygmomanometer (Hawksley-Gelmam, Lancing, Sussex, UK). Pulse pressure was calculated by subtracting the diastolic blood pressure from the systolic blood pressure. Current drug use was based on dispensing labels participants took to their annual visit. As waist circumference was not available in the Hoorn DCS cohort, we estimated this on the basis of age, body mass index, and ethnicity using a previously published method.[40] Cardiovascular events, educational level, ethnic background, and smoking were self-reported.

Models with the top three highest C statistics were validated in GoDARTS, a secondary validation cohort (table 1). GoDARTS is a prospective cohort comprised of people from the Tayside region, Scotland and is described elsewhere.[30] Briefly, laboratory and prescribing data were obtained from medical records and smoking status from a baseline lifestyle questionnaire. Weight and height were measured at recruitment. Outcomes were defined as described in the model development studies (supplementary table S10). Given that albumin was relatively sparsely available in GoDARTS, models with microalbuminuria and macroalbuminuria were not validated, nor were models that included albumin as a predictor. A total of 8698 people from GoDARTS2 and GoDARTS3 with type 2 diabetes were included. The median time to chronic kidney disease from recruitment was 4.7 years (interquartile range 1.98-6.69) and included follow-up data until 2017.

Predicted outcomes

The included outcomes for external validation were microalbuminuria, macroalbuminuria, diabetic kidney disease, and chronic kidney disease. The outcomes defined in our cohort are as described in the individual studies. Generally, diabetic kidney disease and chronic kidney disease were defined as an estimated glomerular filtration rate of <60 mL/min/1.73 m², microalbuminuria as a urinary albumin:creatinine ratio ≥30 mg/g, and macroalbuminuria as an albumin:creatinine ratio ≥300 mg/g. Supplementary table S10 lists the definitions used in each of the studies. All models from the original publications were considered but were excluded if a missing predictor could not be replaced by a proxy or when the frequency of events was too low to allow accurate validation. The latter was the case for the end stage renal disease models, where the incidence was low, with 8, 19, and 28 cases at the 2, 5, and 10 year horizon, respectively.

Statistical analysis

Missing predictor data were imputed five times using the mice function from the R package MICE. Missing values ranged from 1.4% for total cholesterol to 6.9% for urinary albumin; 2868 individuals had one missing variable or more in one of the included predictors. Baseline characteristics were given as means with standard deviations, and as medians with an

interquartile range for not normally distributed data. For each validated model, three prediction horizons of 2, 5, and 10 years were assessed. The performance was estimated based on calibration and discrimination. Calibration was assessed in three ways: visually by plotting the predicted risk against the observed risk in a calibration plot, comparing the observed and expected outcome, and performing the Hosmer-Lemeshow χ^2 test. A model is considered well calibrated when observed and expected outcome is close to 1 and when the observed versus predicted risk fit on the x=y line. Given that most retrieved prediction models did not report an intercept, we recalibrated all models based on the incidence of the Hoorn DCS cohort outcome. Two models could not be recalibrated because the base risk was small, close to zero (macroalbuminuria in the study by Basu et al[41] and albuminuria in the study by Jardine et al[43]) and are reported without recalibration. Discrimination of the models was evaluated using Harrell's C statistic. All analyses were performed in R (version 4.0.2), and figures were produced using *ggplot2*.

Patient and public involvement

Speaking to patient councils inspired this systematic review. Study participants and members of the public were not involved in the design, conduct, or reporting of our research.

Results
Systematic review results
Characteristics of included models
Overall, 10 883 of 11 009 articles screened on title and abstract were excluded, leaving 126 for full text review. When the full text of papers was screened, a total of 87 studies were excluded primarily because the studies did not develop a prediction model or because the study design was cross sectional (supplementary fig S1). Finally, 41 studies were included of which 29 developed one model or more specifically in people with a diagnosis of type 2 diabetes and 12 developed one model or more for use in the general population but with diabetes as predictor (see supplementary file for references of included studies). Overall, these studies accounted for 64 prediction models that predicted some nephropathy outcome, of which 46 were developed in people with diabetes and 18 in the general population (supplementary fig S1). The sample size for models developed in people with diabetes ranged from 86 to 660 856 (supplementary table S4). For the general population models, the sample size ranged from 1549 to 799 658 individuals. The models developed from data of patients with diabetes mainly predicted microalbuminaria, macroalbuminuria, diabetic kidney disease, and end stage renal disease as outcomes (supplementary table S4). General populations models were developed to predict chronic kidney disease and end stage renal disease (supplementary table S6). Although various prediction horizons were used across the models (supplementary table S6), most models used a five year prediction horizon. The most

common predictors for the 42 models developed in people with diabetes were age (22 models), estimated glomerular filtration rate at baseline (21 models), and systolic blood pressure (18 models, supplementary table S5). For the 20 general population models, the most frequently used predictors in addition to the inclusion criterion diabetes were age (14 models) and hypertension (eight models, supplementary table S7).

Risk of bias and applicability
Most of the studies scored a high risk of bias in the analysis domain (supplementary fig S2a-c), mainly as a result of incorrect handling of missing data and lack of adjustment for overfitting (supplementary fig S2). The participants domain—which covers the data source and selection of individuals based on inclusion and exclusion criteria—was rated as low risk of bias in 68% of the models. The domains predictors and outcomes were rated as low risk of bias in 56% and 51% of the models, respectively. In the remaining models these domains were unclearly reported (supplementary fig S2a and b). Of the included studies, 40% scored high concern for applicability to the objectives of the review, as a result of selective sampling of the participants or predictors that are not typically measured in routine diabetes care, including cystatin C and B type natriuretic peptide. Low concern for applicability was scored for 44% of the studies for all domains (supplementary fig S2a and c).

Apparent performance of the models
The discriminative ability of the prediction models was reported through the area under the curve and C statistics in 25 development studies in people with type 2 diabetes. Values ranged from 0.60 (95% confidence interval 0.56 to 0.64) to 0.99 (not available) (fig 1). Across outcomes, apparent performance varied widely. For example, for albuminuria the performance ranged from 0.62 (0.61 to 0.64) to 0.84 (0.80 to 0.88) and for diabetic kidney disease from 0.61 (0.60 to 0.63) to 0.81 (0.79 to 0.84) (fig 1). For models developed for chronic kidney disease in the general population, the performance ranged from 0.70 (not available) to 0.88 (0.87 to 0.88) (fig 2). For end stage renal disease, the performance ranged from 0.60 (0.56 to 0.64) to 0.99 (not available) in people with type 2 diabetes and 0.85 (0.83 to 0.86) to 0.96 (0.95 to 0.97) in the general population.

In nine studies, calibration was evaluated based on the Hosmer-Lemeshow test. One study used the Greenwood-D'Agostino-Nam test, and six studies used a calibration plot, two merely a description of the calibration, and two a calibration table; nine studies did not report on calibration (supplementary table S4). Almost all studies showed good calibration with the Hosmer-Lemeshow test P>0.05. In one study, four models did not pass the Greenwood-D'Agostino-Nam test for calibration.[41]

Eleven studies developed in the general population reported on discrimination, with C statistics ranging from 0.59 (confidence interval not available) to 0.96

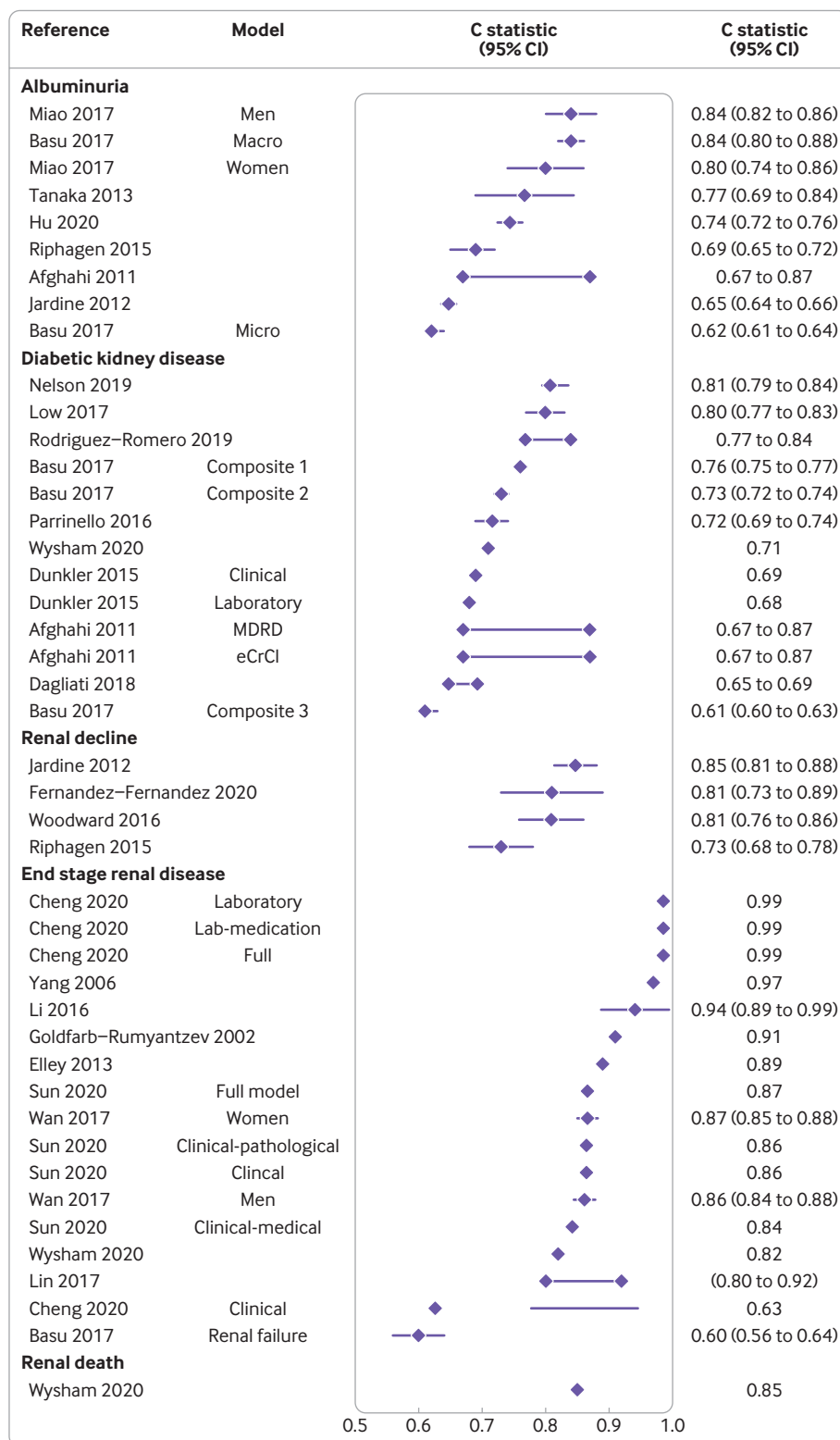| Reference	Model	C statistic (95% CI)	C statistic (95% CI)
Albuminuria			
Miao 2017	Men		0.84 (0.82 to 0.86)
Basu 2017	Macro		0.84 (0.80 to 0.88)
Miao 2017	Women		0.80 (0.74 to 0.86)
Tanaka 2013			0.77 (0.69 to 0.84)
Hu 2020			0.74 (0.72 to 0.76)
Riphagen 2015			0.69 (0.65 to 0.72)
Afghahi 2011			0.67 to 0.87
Jardine 2012			0.65 (0.64 to 0.66)
Basu 2017	Micro		0.62 (0.61 to 0.64)
Diabetic kidney disease			
Nelson 2019			0.81 (0.79 to 0.84)
Low 2017			0.80 (0.77 to 0.83)
Rodriguez–Romero 2019			0.77 to 0.84
Basu 2017	Composite 1		0.76 (0.75 to 0.77)
Basu 2017	Composite 2		0.73 (0.72 to 0.74)
Parrinello 2016			0.72 (0.69 to 0.74)
Wysham 2020			0.71
Dunkler 2015	Clinical		0.69
Dunkler 2015	Laboratory		0.68
Afghahi 2011	MDRD		0.67 to 0.87
Afghahi 2011	eCrCl		0.67 to 0.87
Dagliati 2018			0.65 to 0.69
Basu 2017	Composite 3		0.61 (0.60 to 0.63)
Renal decline			
Jardine 2012			0.85 (0.81 to 0.88)
Fernandez–Fernandez 2020			0.81 (0.73 to 0.89)
Woodward 2016			0.81 (0.76 to 0.86)
Riphagen 2015			0.73 (0.68 to 0.78)
End stage renal disease			
Cheng 2020	Laboratory		0.99
Cheng 2020	Lab-medication		0.99
Cheng 2020	Full		0.99
Yang 2006			0.97
Li 2016			0.94 (0.89 to 0.99)
Goldfarb–Rumyantzev 2002			0.91
Elley 2013			0.89
Sun 2020	Full model		0.87
Wan 2017	Women		0.87 (0.85 to 0.88)
Sun 2020	Clinical-pathological		0.86
Sun 2020	Clincal		0.86
Wan 2017	Men		0.86 (0.84 to 0.88)
Sun 2020	Clinical-medical		0.84
Wysham 2020			0.82
Lin 2017			(0.80 to 0.92)
Cheng 2020	Clinical		0.63
Basu 2017	Renal failure		0.60 (0.56 to 0.64)
Renal death			
Wysham 2020			0.85

0.5 0.6 0.7 0.8 0.9 1.0

Fig 1 | Apparent discrimination for models developed in people with type 2 diabetes across studies and outcomes (see supplementary file for complete reference list of included studies). Whiskers represent 95% confidence intervals or interquartile ranges (Nelson 2019). Two interconnected diamonds represent reported range of C statistics. In Basu 2017 composite outcomes were used. Composite 1=doubling of serum creatinine or ›20 mL/min/1.73 m² decrease in estimated glomerular filtration rate (eGFR); composite 2=macroalbuminuria, renal failure, end stage renal disease, doubling of serum creatinine or ›20 mL/min/1.73 m² decrease in eGFR; composite 3=macroalbuminuria, microalbuminuria, renal failure, or end stage renal disease. MDRD=Modification of Diet in Renal Disease study; eCrCl=estimated creatinine clearance

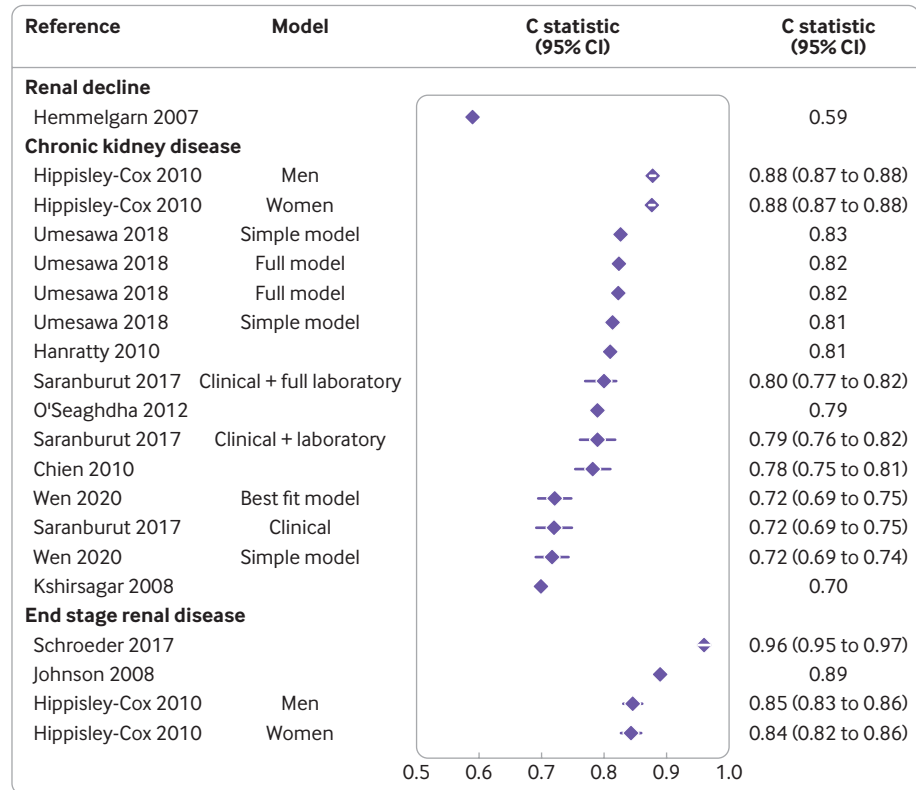| Reference	Model	C statistic (95% CI)	C statistic (95% CI)
Renal decline			
Hemmelgarn 2007			0.59
Chronic kidney disease			
Hippisley-Cox 2010	Men		0.88 (0.87 to 0.88)
Hippisley-Cox 2010	Women		0.88 (0.87 to 0.88)
Umesawa 2018	Simple model		0.83
Umesawa 2018	Full model		0.82
Umesawa 2018	Full model		0.82
Umesawa 2018	Simple model		0.81
Hanratty 2010			0.81
Saranburut 2017	Clinical + full laboratory		0.80 (0.77 to 0.82)
O'Seaghdha 2012			0.79
Saranburut 2017	Clinical + laboratory		0.79 (0.76 to 0.82)
Chien 2010			0.78 (0.75 to 0.81)
Wen 2020	Best fit model		0.72 (0.69 to 0.75)
Saranburut 2017	Clinical		0.72 (0.69 to 0.75)
Wen 2020	Simple model		0.72 (0.69 to 0.74)
Kshirsagar 2008			0.70
End stage renal disease			
Schroeder 2017			0.96 (0.95 to 0.97)
Johnson 2008			0.89
Hippisley-Cox 2010	Men		0.85 (0.83 to 0.86)
Hippisley-Cox 2010	Women		0.84 (0.82 to 0.86)

0.5 0.6 0.7 0.8 0.9 1.0

Fig 2 | Apparent discrimination for models developed in the general population across studies and outcomes (see supplementary file for complete reference list of included studies). Two interconnected diamonds represent reported range of C statistics. Whiskers represent 95% confidence intervals

(0.95 to 0.97), whereas one study did not report discrimination (supplementary table S6). Calibration was assessed in eight studies based on the Hosmer-Lemeshow test and one study based on a calibration plot, one study gave a description of the calibration, and two studies did not report on the calibration. The P values based on the Hosmer-Lemeshow test ranged from 0.01 to 0.99, indicating poor to excellent calibration, and all models but one passed the test (P>0.05).

External validation in Hoorn DCS cohort and GoDARTS

Of the 64 identified models, 21 could be externally validated; 15 developed in a population with type 2 diabetes and six in the general population. Studies were mainly excluded because they predicted end stage renal disease for which the number of events was too low in the Hoorn DCS cohort (26 models), with fewer than 30 events.[42] A second reason for exclusion was that the variables in the prediction models were not available in the Hoorn DCS cohort, such as serum uric acid, cystatin C, and B type natriuretic peptide (16 models). Finally, one model was part of a larger simulation model (supplementary tables S8 and S9). Table 1 shows the baseline characteristics of the the Hoorn DCS cohort. The average age was 62.6 years (SD 12.1 years), and the median duration of diabetes was 0.75 years (intequartile range 0.2-3.8 years). Follow-up ranged from one year to 21.7 years. At the

two year horizon, 1691 people had microalbuminuria, 263 had macroalbuminuria, and 2159 had an eGFR <60 mL/min/1.73 m^2 (supplementary table S10). Top performing models were validated in the secondary cohort, GoDARTS (table 1). The average age in GoDARTS was 65.2 (SD 11.1) years, and the median duration of diabetes was 7.1 (intequartile range 2.9-10.4) years.

Discrimination

The discriminatory ability of the models developed in people with diabetes varied considerably, with C statistics ranging from 0.50 to 0.96 in the Hoorn DCS cohort (fig 3). The performance on the two year horizon was generally best within a model, followed by the five year and 10 year horizons. For most models, however, the difference in performance between the horizons within a model was smaller than the performance between models. In the Hoorn DCS cohort, no apparent difference was found in performance between the models developed in people with diabetes and the general population. In GoDARTS, however, the models developed in the general population performed generally worse than those developed specifically in people with diabetes (fig 3).

The discriminatory ability for albuminuria expressed as C statistics ranged from 0.55 (0.54 to 0.56) to 0.96 (0.95 to 0.97). For macroalbuminuria, two models performed well, with C statistics >0.8, whereas the other models showed poorer performances, with C statistics

Table 1 | Characteristics of external validation cohorts. Values are means (standard deviations) unless stated otherwise

Characteristics	Hoorn DCS cohort	GoDARTS
No	11 450	8698
Age (years)	62.6 (12.1)	65.2 (11.1)
No (%) men	6039 (52.7)	4882 (56)
Median (IQR) diabetes duration (years)	0.75 (0.2-3.8)	7.1 (2.9-10.4)
Body mass index	29.3 (5.4)	31.4 (6.2)
Blood pressure (mm Hg):		
Systolic	141 (20.7)	142 (19)
Diastolic	80 (10.0)	77 (12)
Cholesterol types (mmol/L):		
Total	5.02 (1.1)	4.4 (0.95)
Low density lipoprotein	2.97 (1.0)	2.01 (0.77)
High density lipoprotein	1.22 (0.3)	1.28 (0.39)
Triglyceride	1.89 (1.2)	1.88 (1.3)
Glucose (mmol/L)	8.37 (2.3)	
Median (IQR) HbA$_{1c}$ (mmol/mol)	50.0 (44.3-60.7)	54.1 (46.4-65.03)
HbA$_{1c}$ (%)	6.7 (6.2 to 7.7)	7.1 (6.4-8.1)
No (%) current smoker	2438 (21)	1427 (16.4)
No (%) of European descent	10 372 (90.6)	8698 (100)
No (%) using antihypertensives	6592 (57.6)	7075 (82)
No (%) using lipid lowering drugs	4697 (41)	7781 (90)
No (%) using glucose lowering drugs:		
None	3349 (29.2)	2671 (30.7)
Oral	6893 (60.2)	4515 (52.0)
Oral+insulin	574 (5.0)	377 (4.3)
Insulin	634 (5.5)	1135 (13.0)

DCS=Diabetes Care System; GoDARTS=Genetics of Diabetes Audit and Research in Tayside Scotland; IQR=interquartile range.

<0.7. The best performing model was that of Basu et al for macroalbuminuria, with a C statistic of 0.96 (0.95 to 0.97) at the two year horizon (fig 3).[41] The model also performed well at the five and 10 year horizons, with C statistics of 0.91 (0.90 to 0.92) and 0.87 (0.85 to 0.88), respectively. The second best performing albuminuria model was of Jardine et al, with C statistics of 0.86 (0.85 to 0.87), 0.84 (0.83 to 0.85), and 0.78 (0.77 to 0.79) at the 2, 5, and 10 year horizon, respectively (fig 3).[43] Afghahi et al's model for albuminuria performed the worst, with C statistics ranging from 0.55 to 0.59 for the three horizons (fig 3).[44]

Four models showed good performance for diabetic kidney disease models, with C statistics >0.75, whereas the other six models showed poorer performances, with C statistics <0.7. Afghahi et al's model performed best, especially when the eGFR was based on the Cockcroft-Gault equation, with C statistics of 0.95 (0.95 to 0.95), 0.93 (0.92 to 0.93), and 0.91 (0.91 to 0.91) at the 2, 5, and 10 year horizon, respectively (fig 3).[44] The second best model was that of Nelson et al, which was based on eGFR from the Modification of Diet in Renal Disease study, with C statistics of 0.87 (0.87 to 0.88), 0.81 (0.81 to 0.82), and 0.76 (0.75 to 0.77) at the 2, 5, and 10 year horizon, respectively.[28] Dagliati et al's model performed worst, with C statistics of 0.51 (0.50 to 0.51), 0.52 (0.51 to 0.52), and 0.50 (0.49 to 0.51) at the 2, 5, and 10 year horizon, respectively.[45] Observed C statistics in the current study's data were similar to the external validation performed in the development study, with 0.81 at the five year horizon (supplementary table S11). In addition, the C statistic of Dunkler et al's[46] model was similar in the current

study compared with the performance observed by Dunkler in their discovery and validation at the five year horizon, with C statistics ranging from 0.68 to 0.70 (supplementary table S11). Good performances were observed for Afghahi et al's model based on eGFR from the Modification of Diet in Renal Disease study, with C statistics of 0.83 (0.83 to 0.84), 0.80 (0.80 to 0.81), and 0.78 (0.77 to 0.79) at the 2, 5 and 10 year horizon, respectively.[44] Of the top three best performing models, Afghahi et al's two models could be validated in GoDARTS (fig 3).[44] The discrimination in GoDARTS outperformed that in the Hoorn DCS cohort, with C statistics of 0.99 (0.99 to 0.99) and 0.85 (0.84 to 0.86) for the endpoint based on estimated creatine clearance and the Modification of Diet in Renal Disease study, respectively.

Of the six models for chronic kidney disease developed in the general population, a clear distinction was apparent between model performance. The C statistics for the performance of the top three models were >0.8, whereas the remaining models performed noticeably worse. Hanratty et al's model for chronic kidney disease performed best, with C statistics of 0.94 (0.94 to 0.94), 0.90 (0.89 to 0.90), and 0.85 (0.85 to 0.86) at the 2, 5, and 10 year horizon, respectively.[47] Saranburut et al's laboratory model also performed well, with C statistics at the corresponding three horizons of 0.89 (0.89 to 0.90), 0.86 (0.86 to 0.87), and 0.84 (0.83 to 0.85).[48] The worst performing model was that of Wen et al, with C statistics ranging from 0.57 to 0.59.[49] The top three best performing models were additionally validated in GoDARTS, where the discrimination was considerably lower than in the Hoorn DCS cohort (fig 3), with C statistics of 0.74 (0.73 to 0.75) in Hanratty et al's model,[47] 0.75 (0.74 to 0.76) in Saranburut et al's model,[48] and 0.71 (0.69 to 0.73) in O'Seaghdha et al's model.[50] The models of O'Seaghdha et al[50] and Chien et al[51] have been externally validated in both Nelson et al[28] and in Fraccaro et al[26] (supplementary table S11). In Nelson et al's study,[28] the discrimination was similar to the C statistics observed in the current study at the five year horizon, with a C statistic 0.81. The C statistics >0.90 reported in Fraccaro et al[26] for both Chien et al's and O'Seaghdha et al's models were much higher than those observed in the Hoorn DCS cohort and GoDARTS and in the external validation by Nelson et al[28] (supplementary table S11).

When the performance of the models was investigated between men and women (supplementary fig S3), a high correlation was found between the C statistics of models in men and women (Pearson's r ≥0.96), suggestive of almost identical performance of investigated models in both sexes.

Calibration

The calibration of models was generally good after recalibration based on the incidence of the specific outcome in the Hoorn DCS cohort. Two models are reported without recalibration (Basu,[41] for macroalbuminuria and Jardine,[43] for albuminuria).
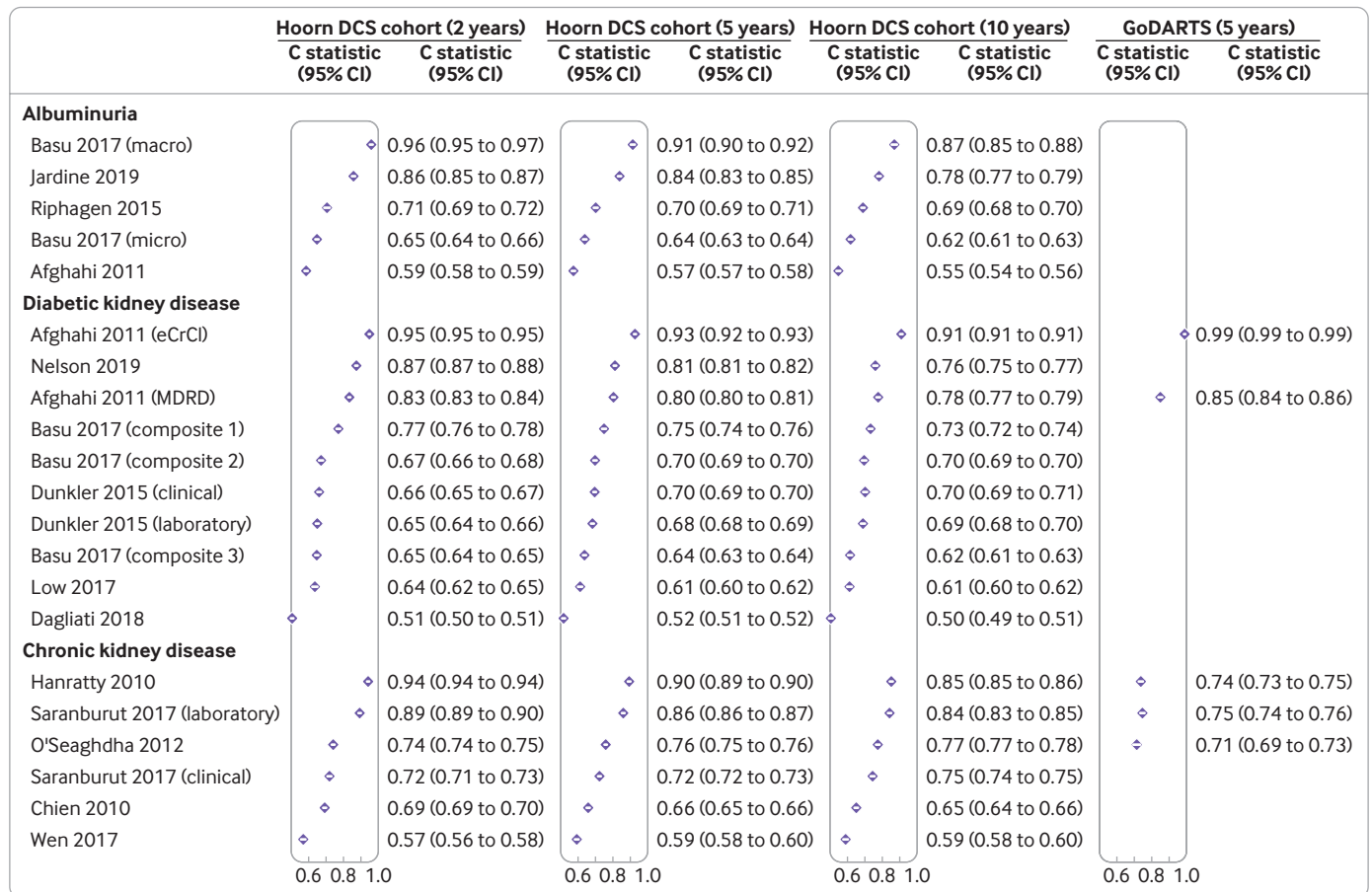
	Hoorn DCS cohort (2 years)		Hoorn DCS cohort (5 years)		Hoorn DCS cohort (10 years)		GoDARTS (5 years)	
	C statistic (95% CI)	C statistic (95% CI)	C statistic (95% CI)	C statistic (95% CI)	C statistic (95% CI)	C statistic (95% CI)	C statistic (95% CI)	C statistic (95% CI)
Albuminuria								
Basu 2017 (macro)		0.96 (0.95 to 0.97)		0.91 (0.90 to 0.92)		0.87 (0.85 to 0.88)		
Jardine 2019		0.86 (0.85 to 0.87)		0.84 (0.83 to 0.85)		0.78 (0.77 to 0.79)		
Riphagen 2015		0.71 (0.69 to 0.72)		0.70 (0.69 to 0.71)		0.69 (0.68 to 0.70)		
Basu 2017 (micro)		0.65 (0.64 to 0.66)		0.64 (0.63 to 0.64)		0.62 (0.61 to 0.63)		
Afghahi 2011		0.59 (0.58 to 0.59)		0.57 (0.57 to 0.58)		0.55 (0.54 to 0.56)		
Diabetic kidney disease								
Afghahi 2011 (eCrCl)		0.95 (0.95 to 0.95)		0.93 (0.92 to 0.93)		0.91 (0.91 to 0.91)		0.99 (0.99 to 0.99)
Nelson 2019		0.87 (0.87 to 0.88)		0.81 (0.81 to 0.82)		0.76 (0.75 to 0.77)		
Afghahi 2011 (MDRD)		0.83 (0.83 to 0.84)		0.80 (0.80 to 0.81)		0.78 (0.77 to 0.79)		0.85 (0.84 to 0.86)
Basu 2017 (composite 1)		0.77 (0.76 to 0.78)		0.75 (0.74 to 0.76)		0.73 (0.72 to 0.74)		
Basu 2017 (composite 2)		0.67 (0.66 to 0.68)		0.70 (0.69 to 0.70)		0.70 (0.69 to 0.70)		
Dunkler 2015 (clinical)		0.66 (0.65 to 0.67)		0.70 (0.69 to 0.70)		0.70 (0.69 to 0.71)		
Dunkler 2015 (laboratory)		0.65 (0.64 to 0.66)		0.68 (0.68 to 0.69)		0.69 (0.68 to 0.70)		
Basu 2017 (composite 3)		0.65 (0.64 to 0.65)		0.64 (0.63 to 0.64)		0.62 (0.61 to 0.63)		
Low 2017		0.64 (0.62 to 0.65)		0.61 (0.60 to 0.62)		0.61 (0.60 to 0.62)		
Dagliati 2018		0.51 (0.50 to 0.51)		0.52 (0.51 to 0.52)		0.50 (0.49 to 0.51)		
Chronic kidney disease								
Hanratty 2010		0.94 (0.94 to 0.94)		0.90 (0.89 to 0.90)		0.85 (0.85 to 0.86)		0.74 (0.73 to 0.75)
Saranburut 2017 (laboratory)		0.89 (0.89 to 0.90)		0.86 (0.86 to 0.87)		0.84 (0.83 to 0.85)		0.75 (0.74 to 0.76)
O'Seaghdha 2012		0.74 (0.74 to 0.75)		0.76 (0.75 to 0.76)		0.77 (0.77 to 0.78)		0.71 (0.69 to 0.73)
Saranburut 2017 (clinical)		0.72 (0.71 to 0.73)		0.72 (0.72 to 0.73)		0.75 (0.74 to 0.75)		
Chien 2010		0.69 (0.69 to 0.70)		0.66 (0.65 to 0.66)		0.65 (0.64 to 0.66)		
Wen 2017		0.57 (0.56 to 0.58)		0.59 (0.58 to 0.60)		0.59 (0.58 to 0.60)		
	0.6 0.8 1.0		0.6 0.8 1.0		0.6 0.8 1.0		0.6 0.8 1.0	

Fig 3 | Discriminatory ability of models included in external validation by outcome and prediction horizons. Whiskers represent 95% confidence intervals. See supplementary file for complete reference list of included studies. In Basu 2017 three composite outcomes were defined: composite 1=doubling of serum creatinine or >20 mL/min/1.73m^2 decrease in estimated glomerular filtration rate (eGFR); composite 2=macroalbuminuria, renal failure, end stage renal disease, and doubling of serum creatinine or >20 mL/min/1.73m^2 decrease in eGFR; composite 3=macroalbuminuria, microalbuminuria, renal failure, or end stage renal disease. DCS=Diabetes Care System; MDRD=Modification of Diet in Renal Disease study; eCrCl=estimated creatinine clearance

Two of the five models developed for albuminuria (Basu,[41] Jardine,[43] supplementary fig S4) showed an overestimation of the risk, whereas the other three models showed good calibration. No differences were observed between horizons. All albuminuria models were significant in the Hosmer-Lemeshow χ^2 test, except for Basu et al's microalbuminuria model.[41] The models developed for diabetic kidney disease developed well in terms of observed to expected ratios, with ratios ≥0.99 in nine models. The three models of Basu et al[41] and two models of Dunkler et al[46] showed the best calibration (supplementary fig S5h). The poorest calibration was observed for the models of Low et al[52] and Dagliati et al[45] (supplementary fig S5). Again, the model of Basu et al showed the best calibration (supplementary fig S5h). The observed to expected ratios were generally close to 1, with Nelson et al's model having the highest deviation.[28]

Five of the six general population models showed good calibration based on observed to expected ratios. Calibration was best for Saranburut et al's clinical model.[48] Chien et al's model showed an overestimation of risk[51] (supplementary fig S6).

Discussion

We conducted a systematic review and quality assessment of prediction models for nephropathy in people with type 2 diabetes. Overall, 46 models were developed in a type 2 diabetes population and 18 in the general population. These studies accounted for a total of 64 prediction models for albuminuria, renal impairment, diabetic kidney disease, chronic kidney disease, and end stage renal disease. Of the 64 prediction models, 21 were externally validated in the Hoorn Diabetes Care System cohort. The reported discrimination of the models varied considerably across outcomes and models and, to some extent, prediction horizon. Multiple models performed well, with C statistics >0.80 for the three investigated outcomes of albuminuria, chronic kidney disease, and diabetic kidney disease. The calibration showed the same variation between outcomes and studies compared with the discrimination, but to a lesser extent between horizons. The models performed better in terms of discrimination at the two year horizon compared with the 5 and 10 year horizons. For most models, however, the difference in performance

between the horizons within a model was smaller than the performance between models. Models developed in people with diabetes did, in general, perform similar to models developed in the general population.

Principal findings

Of the five validated models for albuminuria, those by Basu et al[41] and Jardine et al[43] performed best in terms of discrimination. Basu et al's models showed better calibration than Jardine et al's models. Jardine et al's model has eight predictors compared with 14 in Basu et al's model. In terms of clinical practicality, both models are equal and contain routinely available predictors. Basu et al's model, however, has the highest number of predictors, including less common accessible predictors, such as cardiovascular disease and the use of anticoagulants.

Ten models for diabetic kidney disease were validated. In terms of discrimination, the two models of Afghahi et al[44] and the model of Nelson et al[28] performed best. The estimated creatinine clearance of the model of Afghahi et al[44] uses six and the model of Nelson et al[28] uses 12 routinely available predictors. In terms of calibration, the model of Afghahi et al outperformed that of Nelson et al.[28] Models with fewer variables, such as the those of Afghahi et al[44] and Saranburut et al,[48] especially in resource poor settings, might be more useful than models with many variables, such as Nelson et al's model[28] (supplementary table S12). Furthermore, both models of Afghahi et al[44] also showed excellent discrimination in GoDARTS.

For the models developed for chronic kidney disease in the general population, Hanratty et al's model[47] and Saranburut et al's laboratory model[48] performed well in a population with type 2 diabetes. Saranburut et al's model, however, showed a better calibration than Hanratty et al's model. Both models contain routinely available predictors, with five (age, diabetes, peripheral vascular disease, eGFR, period of observation) predictors in Hanratty et al's model and six predictors (age, sex, systolic blood pressure, waist circumference, diabetes, eGFR) in Saranburut et al's model, and could therefore be applied in clinical practice (supplementary table S12). The discrimination of Hanratty et al's and Saranburut et al's models was lower in GoDARTS.

Implications

As the number of people with type 2 diabetes increases worldwide, the incidence of diabetic kidney disease will also increase. It is therefore vital to identify those at higher risk of nephropathy to be able to enhance monitoring and possibly to intervene at an early stage to slow down renal decline. Interventions include intensified glycaemic control; nephroprotective glucose lowering drugs, including glucagon-like peptide 1 receptor antagonists and sodium-glucose cotransporter 2 inhibitors; blood pressure control; and renin-angiotensin system inhibition.[14-17] More stringent glycaemic control has been suggested to reduce the development and worsening of

nephropathy.[13] Prediction models will help to estimate the risk of diabetic kidney disease, especially when the models are based on routine clinical markers. Our results show that for the endpoints albuminuria, diabetic kidney disease, and chronic kidney disease, prediction models are available with the capability of reliably discriminating between people with a low and a high risk. In the validation study, models that predict chronic kidney disease and diabetic kidney disease (eGFR <60 mL/min/1.73m^2) performed well across all three prediction horizons, which could give people with diabetes an estimation of risk for the near future (two year risk) and in the long term (10 year). For example, Saranburut et al's model[48] and Afghahi et al's model[44] both showed good discrimination and validation. However, we observed that the two year horizon outperformed the 5 and 10 year horizons, which is not unexpected given that baseline measures are more likely to predict the near future accurately. In contrast, in the long term, much can change in terms of, for example, changes in lifestyle or treatment regimens and thus outcome events.

Strengths and limitations of this study

This systemic review has some limitations and strengths. One limitation is that we excluded articles that were not in English or Dutch. During full text screening, although we excluded three studies based on language,[44-46] we were still able to review the reference lists for additional eligible studies. Other partially comparable systematic reviews did not include any of these three studies as relevant prediction models.[23 25 53] Therefore, it can be questioned whether the excluded studies were relevant to this systematic review. A second limitation is that we could not validate models that predicted end stage renal disease because of too limited events in our study population; more investigation is needed. The options to intervene in end stage renal disease are limited, and as such interventions ideally should occur earlier. A third limitation is that the Hoorn DCS cohort comprised people with generally well controlled diabetes. As such, models developed on high risk populations performed less well in the external validation in contrast with models developed on lower risk populations. Finally, we were unable to use several models because comparable measurements were not available in the Hoorn DCS cohort. However, given that these variables were not measured as part of routine care is an indication of their limited applicability.

The study's strengths include the systematic review to identify relevant prognostic models and quality assessment of the included studies. Secondly, we externally validated the models in a large prospective cohort study with long term follow-up and detailed phenotyping based on routine care data thus enabling a head-to-head comparison of the models. Thirdly, we validated models with the best discrimination in a secondary cohort from a different country. Future studies should evaluate the generalisability of the models in people with other forms of diabetes and to

what extent these models have been implemented in day to day clinical practice.

Conclusion

Many prognostic models are available to predict nephropathy in people with type 2 diabetes. We identified 64 prediction models, 21 of which could be validated and directly compared in the Hoorn DCS cohort. For each of the three included outcomes of albuminuria, diabetic kidney disease, and chronic kidney disease, discrimination and calibration varied considerably across horizons and models. Several models showed good discrimination and calibration across various prediction horizons for each outcome, although models performed best at the 2 year horizon. In a secondary validation cohort, models developed for diabetic kidney disease especially showed good discrimination compared with those developed for chronic kidney disease. This study identified several suitable models that will contribute to preventing or postponing renal decline and ultimately end stage renal disease in people with diabetes.

AUTHOR AFFILIATIONS

[1]Department of Epidemiology and Data Science, Amsterdam Public Health Institute, Amsterdam Cardiovascular Sciences Institute, Amsterdam UMC, Location VUmc, 1081 HV, Amsterdam, Netherlands

[2]Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, Netherlands

[3]Department of General Practice, Amsterdam Public Health Institute, Amsterdam UMC, Location VUmc, Amsterdam, Netherlands

[4]Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK

[5]PHARMO Institute for Drug Outcomes Research, Utrecht, Netherlands

[6]Groningen Research Institute of Pharmacy, University of Groningen, Groningen, Netherlands

[7]Centre for Nutrition, Prevention and Health Services, Institute for Public Health and the Environment, Bilthoven, Netherlands

[8]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[9]Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[10]Molecular Epidemiology section, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

1   Federation ID. IDF Diabetes Atlas. 2015.
2   van Dieren S, Beulens JW, van der Schouw YT, Grobbee DE, Neal B. The global burden of diabetes and its complications: an emerging pandemic. Eur J Cardiovasc Prev Rehabil 2010;17(Suppl 1):S3-8. doi:10.1097/01.hjr.0000368191.86614.5a
3   Adler AI, Stevens RJ, Manley SE, Bilous RW, Cull CA, Holman RR, UKPDS GROUP. Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). Kidney Int 2003;63:225-32. doi:10.1046/j.1523-1755.2003.00712.x
4   Thomas S, Karalliede J. Diabetic nephropathy. Medicine (Baltimore) 2014;43:20-5. doi:10.1016/j.mpmed.2014.10.007
5   de Boer IH, Rue TC, Hall YN, Heagerty PJ, Weiss NS, Himmelfarb J. Temporal trends in the prevalence of diabetic kidney disease in the United States. JAMA 2011;305:2532-9. doi:10.1001/jama.2011.861
6   Joshi VD. Quality of life in end stage renal disease patients. World J Nephrol 2014;3:308-16. doi:10.5527/wjn.v3.i4.308
7   Zoccali C, Kramer A, Jager KJ. Chronic kidney disease and end-stage renal disease-a review produced to contribute to the report 'the status of health in the European union: towards a healthier Europe'. NDT Plus 2010;3:213-24.
8   Ayodele OE, Alebiosu CO, Salako BL. Diabetic nephropathy--a review of the natural history, burden, risk factors and treatment. J Natl Med Assoc 2004;96:1445-54.
9   Keane WF, Brenner BM, de Zeeuw D, et al, RENAAL Study Investigators. The risk of developing end-stage renal disease in patients with type 2 diabetes and nephropathy: the RENAAL study. Kidney Int 2003;63:1499-507. doi:10.1046/j.1523-1755.2003.00885.x
10 Harjutsalo V, Groop P-H. Epidemiology and risk factors for diabetic kidney disease. Adv Chronic Kidney Dis 2014;21:260-6. doi:10.1053/j.ackd.2014.03.009
11 Curtis S, Komenda P. Screening for chronic kidney disease: moving toward more sustainable health care. Curr Opin Nephrol Hypertens 2020;29:333-8. doi:10.1097/MNH.0000000000000597
12 Brocco E, Fioretto P, Mauer M, et al. Renal structure and function in non-insulin dependent diabetic patients with microalbuminuria. Kidney Int Suppl 1997;63:S40-4.
13 Sardar P, Udell JA, Chatterjee S, Bansilal S, Mukherjee D, Farkouh ME. Effect of Intensive Versus Standard Blood Glucose Control in Patients With Type 2 Diabetes Mellitus in Different Regions of the World: Systematic Review and Meta-analysis of Randomized Controlled Trials. J Am Heart Assoc 2015;4:e001577. doi:10.1161/JAHA.114.001577

14 Umesawa M, Sairenchi T, Haruyama Y, et al. Validity of a Risk Prediction Equation for CKD After 10 Years of Follow-up in a Japanese Population: The Ibaraki Prefectural Health Study. *Am J Kidney Dis* 2018;71:842-50. doi:10.1053/j.ajkd.2017.09.013

15 Piperidou A, Loutradis C, Sarafidis P. SGLT-2 inhibitors and nephroprotection: current evidence and future perspectives. *J Hum Hypertens* 2021;35:12-25. doi:10.1038/s41371-020-00393-4

16 Piperidou A, Sarafidis P, Boutou A, et al. The effect of SGLT-2 inhibitors on albuminuria and proteinuria in diabetes mellitus: a systematic review and meta-analysis of randomized controlled trials. *J Hypertens* 2019;37:1334-43. doi:10.1097/HJH.0000000000002050

17 Mosterd CM, Bjornstad P, van Raalte DH. Nephroprotective effects of GLP-1 receptor agonists: where do we stand?*J Nephrol* 2020;33:965-75. doi:10.1007/s40620-020-00738-9

18 DeFronzo RA, Norton L, Abdul-Ghani M. Renal, metabolic and cardiovascular considerations of SGLT2 inhibition. *Nat Rev Nephrol* 2017;13:11-26. doi:10.1038/nrneph.2016.170

19 Levey AS, Coresh J, Balk E, et al, National Kidney Foundation. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Ann Intern Med* 2003;139:137-47. doi:10.7326/0003-4819-139-2-200307150-00013

20 Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98:691-8. doi:10.1136/heartjnl-2011-301247

21 Collins GS, Moons KG. Comparing risk prediction models. *BMJ* 2012;344:e3186. doi:10.1136/bmj.e3186

22 Tziomalos K, Athyros VG. Diabetic Nephropathy: New Risk Factors and Improvements in Diagnosis. *Rev Diabet Stud* 2015;12:110-8. doi:10.1900/RDS.2015.12.110

23 Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268-77. doi:10.1016/j.jclinepi.2012.06.020

24 van Rijn MHC, van de Luijtgaarden M, van Zuilen AD, et al. Prognostic models for chronic kidney disease: a systematic review and external validation. *Nephrol Dial Transplant* 2020;gfaa155. doi:10.1093/ndt/gfaa155

25 Tangri N, Kitsios GD, Inker LA, et al. Risk prediction models for patients with chronic kidney disease: a systematic review. *Ann Intern Med* 2013;158:596-603. doi:10.7326/0003-4819-158-8-201304160-00004

26 Fraccaro P, van der Veer S, Brown B, et al. An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK. *BMC Med* 2016;14:104. doi:10.1186/s12916-016-0650-2

27 Ramspek CL, de Jong Y, Dekker FW, van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrol Dial Transplant* 2020;35:1527-38. doi:10.1093/ndt/gfz018

28 Nelson RG, Grams ME, Ballew SH, et al, CKD Prognosis Consortium. Development of Risk Prediction Equations for Incident Chronic Kidney Disease. *JAMA* 2019;322:2104-14. doi:10.1001/jama.2019.17379

29 van der Heijden AA, Rauh SP, Dekker JM, et al. The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. *BMJ Open* 2017;7:e015599. doi:10.1136/bmjopen-2016-015599

30 Hébert HL, Shepherd B, Milburn K, et al. Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). *Int J Epidemiol* 2018;47:380-381j. doi:10.1093/ije/dyx140

31 Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460. doi:10.1136/bmj.i6460

32 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097. doi:10.1371/journal.pmed.1000097

33 Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009;62:1013-20. doi:10.1016/j.jclinepi.2008.10.009

34 Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. doi:10.7326/M14-0698

35 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594. doi:10.1136/bmj.g7594

36 Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744. doi:10.1371/journal.pmed.1001744

37 Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019;170:W1-33. doi:10.7326/M18-1377

38 Wolff RF, Moons KGM, Riley RD, et al, PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170:51-8. doi:10.7326/M18-1376

39 Aldington SJ, Kohner EM, Meuer S, Klein R, Sjølie AK. Methodology for retinal photography and assessment of diabetic retinopathy: the EURODIAB IDDM complications study. *Diabetologia* 1995;38:437-44. doi:10.1007/BF00410281

40 Bozeman SR, Hoaglin DC, Burton TM, Pashos CL, Ben-Joseph RH, Hollenbeak CS. Predicting waist circumference from body mass index. *BMC Med Res Methodol* 2012;12:115. doi:10.1186/1471-2288-12-115

41 Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomised trials. *Lancet Diabetes Endocrinol* 2017;5:788-98. doi:10.1016/S2213-8587(17)30221-8

42 Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-83. doi:10.1016/j.jclinepi.2004.06.017

43 Jardine MJ, Hata J, Woodward M, et al, ADVANCE Collaborative Group. Prediction of kidney-related outcomes in patients with type 2 diabetes. *Am J Kidney Dis* 2012;60:770-8. doi:10.1053/j.ajkd.2012.04.025

44 Afghahi H, Cederholm J, Eliasson B, et al. Risk factors for the development of albuminuria and renal impairment in type 2 diabetes--the Swedish National Diabetes Register (NDR). *Nephrol Dial Transplant* 2011;26:1236-43. doi:10.1093/ndt/gfq535

45 Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol* 2018;12:295-302. doi:10.1177/1932296817706375

46 Dunkler D, Gao P, Lee SF, et al, ONTARGET and ORIGIN Investigators. Risk prediction for early CKD in type 2 diabetes. *Clin J Am Soc Nephrol* 2015;10:1371-9. doi:10.2215/CJN.10321014

47 Hanratty R, Chonchol M, Miriam Dickinson L, et al. Incident chronic kidney disease and the rate of kidney function decline in individuals with hypertension. *Nephrol Dial Transplant* 2010;25:801-7. doi:10.1093/ndt/gfp534

48 Saranburut K, Vathesatogkit P, Thongmung N, et al. Risk scores to predict decreased glomerular filtration rate at 10 years in an Asian general population. *BMC Nephrol* 2017;18:240. doi:10.1186/s12882-017-0653-z

49 Wen J, Hao J, Zhang Y, et al. Risk scores for predicting incident chronic kidney disease among rural Chinese people: a village-based cohort study. *BMC Nephrol* 2020;21:120. doi:10.1186/s12882-020-01787-9

50 O'Seaghdha CM, Lyass A, Massaro JM, et al. A risk score for chronic kidney disease in the general population. *Am J Med* 2012;125:270-7. doi:10.1016/j.amjmed.2011.09.009

51 Chien KL, Lin HJ, Lee BC, Hsu HC, Lee YT, Chen MF. A prediction model for the risk of incident chronic kidney disease. *Am J Med* 2010;123:836-846.e2. doi:10.1016/j.amjmed.2010.05.010

52 Low S, Lim SC, Zhang X, et al. Development and validation of a predictive model for Chronic Kidney Disease progression in Type 2 Diabetes Mellitus based on a 13-year study in Singapore. *Diabetes Res Clin Pract* 2017;123:49-54. doi:10.1016/j.diabres.2016.11.008

53 Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med* 2012;9:e1001344. doi:10.1371/journal.pmed.1001344

Supplementary information: additional tables S1-S12 and references of included studies

Supplementary information: additional figures S1-S6