Gregorich et al. Diagnostic and Prognostic Research (2021) 5:19
https://doi.org/10.1186/s41512-021-00107-5

Diagnostic and
Prognostic Research

PROTOCOL Open Access

A prediction model for the decline in renal function in people with type 2 diabetes mellitus: study protocol

Mariella Gregorich[1,2], Andreas Heinzel[2], Michael Kammer[1,2], Heike Meiselbach[3], Carsten Böger[4], Kai-Uwe Eckardt[5], Gert Mayer[6], Georg Heinze[1] and Rainer Oberbauer[2*]

Abstract

Background: Chronic kidney disease (CKD) is a well-established complication in people with diabetes mellitus. Roughly one quarter of prevalent patients with diabetes exhibit a CKD stage of 3 or higher and the individual course of progression is highly variable. Therefore, there is a clear need to identify patients at high risk for fast progression and the implementation of preventative strategies. Existing prediction models of renal function decline, however, aim to assess the risk by artificially grouped patients prior to model building into risk strata defined by the categorization of the least-squares slope through the longitudinally fluctuating eGFR values, resulting in a loss of predictive precision and accuracy.

Methods: This study protocol describes the development and validation of a prediction model for the longitudinal progression of renal function decline in Caucasian patients with type 2 diabetes mellitus (DM2). For development and internal-external validation, two prospective multicenter observational studies will be used (PROVALID and GCKD). The estimated glomerular filtration rate (eGFR) obtained at baseline and at all planned follow-up visits will be the longitudinal outcome. Demographics, clinical information and laboratory measurements available at a baseline visit will be used as predictors in addition to random country-specific intercepts to account for the clustered data. A multivariable mixed-effects model including the main effects of the clinical variables and their interactions with time will be fitted. In application, this model can be used to obtain personalized predictions of an eGFR trajectory conditional on baseline eGFR values. The final model will then undergo external validation using a third prospective cohort (DIACORE). The final prediction model will be made publicly available through the implementation of an *R shiny* web application.

Discussion: Our proposed state-of-the-art methodology will be developed using multiple multicentre study cohorts of people with DM2 in various CKD stages at baseline, who have received modern therapeutic treatment strategies of diabetic kidney disease in contrast to previous models. Hence, we anticipate that the multivariable prediction model will aid as an additional informative tool to determine the patient-specific progression of renal function and provide a useful guide to early on identify individuals with DM2 at high risk for rapid progression.

Keywords: Prediction modeling, Type 2 diabetes mellitus, Chronic kidney disease, Mixed model, Prognosis, Risk calculator

* Correspondence: rainer.oberbauer@meduniwien.ac.at
[2]Division of Nephrology and Dialysis, Department of Internal Medicine III, Medical University of Vienna, Vienna, Austria
Full list of author information is available at the end of the article

Introduction

Chronic kidney disease (CKD) has become an increasing global health problem, partly due to the continuously rising incidences of diabetes mellitus, obesity, and hypertension [1]. It is estimated that 8 to 16% of the world's population suffer from CKD, and the rate is even higher in people with DM2, where cross-sectional studies report percentages of roughly 25% [2]. Ninety-five percent of people with early CKD are unaware of their disease [3]. However, early prediction of the continuous decline in kidney function could provide an additional resource for personalized preventative care [4]. Personalized risk assessment based on large study cohorts could therefore offer several potential benefits to the preventive care for people with DM2 such as early detection followed by monitoring to guide treatment and potentially slow the progression of the decline of kidney function as expressed by the estimated glomerular filtration rate (eGFR) [4]. Despite the accumulated knowledge with regard to CKD nowadays, identifying individuals at high risk for fast disease progression has proven to be challenging because longitudinal eGFR slopes not only vary between patients (inter-patient variability), but eGFR measurements also fluctuate within each patient over time (intra-patient variability) [5].

Nevertheless, several prediction models for the progression of renal function decline have previously been developed in various populations. It is of note that the models which were developed in studies from the last decade may no longer be applicable nowadays, because the new treatment classes of SGLT2-blocker and mineralocorticoid receptor antagonists have dramatically changed the course of CKD progression [6–8]. In addition, existing predictive models have mainly considered hard renal endpoints such as incidence or end-stage renal failure [9, 10]. Studies focused on renal function have mainly considered simplifications of the longitudinal eGFR trajectory by dichotomization of the expected eGFR slope. For instance, Subasi et al. [11] conducted a pilot study based on a randomized double-blinded treatment trial, the African American Study of Kidney Disease and Hypertension (AASK), in order to predict the rate of kidney function decline. The least-squares slopes of GFR decline from the 6-month time point until censoring were used to define rapid (i.e., an absolute slope < 3mL/min/1.73m^2/year) and slow (i.e., absolute slope in between 1 and 3mL/min/1.73m^2/year) progressors for this study. Similarly, Vigil et al. [12] aimed towards the identification of predictors of a rapid decline of renal function defined as an annual eGFR loss > 4 mL/min/1.73m^2. However, covariates considered for selection in this analysis were chosen on the basis of their significance in univariable analysis (or by their clinical or biological relevance). Despite its popular use for data reduction, univariable prefiltering has been shown to lead to suboptimal or biased selection results [13]. Pena et al. [14] showed that a panel of novel biomarkers improves the prediction of renal function decline in people with DM2. In contrast to the other mentioned studies, renal function decline was not dichotomized. The eGFR slope was estimated using linear regression and then constituted the outcome of interest in its continuous form. People with an observed eGFR decline larger than −3mL/min/1.73m^2 were considered stable, all others were considered to be rapid progressors in terms of kidney function decline. Then, a person's risk was assessed by dichotomizing the observed decline in these two groups and comparing it with the predicted probabilities of eGFR decline. The threshold of −3mL/min/1.73m^2 was selected based on literature, which however included older studies without the current state-of-the-art therapy of diabetic kidney disease, where a GFR loss of only 1mL/min//1.73m^2/year is standard [14–17]. Furthermore, molecular data in regular clinical care is usually not easily available and no validated reference values exist for these markers.

In this study, we will apply state-of-the-art methodology to avoid deficiencies of previous attempts in prediction models for renal function decline. We will directly incorporate the repeated eGFR measurements per person as an outcome vector into a multivariable mixed-effects model accounting for the dependence of the repeated measurements of eGFR and the clustered data structure due to the use of two multicenter prospective study cohorts. We will also provide the methodology to condition a prediction on an available baseline value of eGFR, facilitating predictions regardless of whether a baseline eGFR is available or not. In contrast to the commonly applied approaches, the repeated values of eGFR across follow-up visits will not be used for slope estimation followed by categorization to infer groups of the severity of loss in kidney function (e.g., stable, mild, and rapid progression) prior to model development. The proposed approach will enable the identification of groups of patients with a high risk of rapid renal function decline by longitudinal prediction of subject-specific eGFR trajectories, thereby providing the probability of exceeding a prespecified cutpoint of eGFR decline per year. After the planned external validation, our model will be implemented as a web tool and can be clinically applied to identify people with DM2 at increased risk of rapid decline in kidney function, or as a supporting tool for clinicians in medical decision-making so that patients can be prompted to seek medical care before a significant deterioration in kidney function occurs.

Objective

The main objective of the planned analysis is the development, internal-external, and external validation of a personalized prediction model for eGFR loss per year in Caucasian people with DM2. It is based on clinical information and laboratory measurements recorded at baseline. A secondary aim is the implementation of the externally validated prediction model as a publicly available web application to facilitate general applicability in clinical care based on data from recent eras using modern therapy.

Research design and methods

Data source

The study cohort for a model building comprises individuals from two distinct prospective observational studies (PROVALID, GCKD) covering a wide range of CKD states [18, 19]. For the purpose of external validation, a third prospective cohort study (DIACORE) will be used [20].

PROVALID - PROspective cohort study in patients with type 2 diabetes mellitus for VALIDation of biomarkers

The PROVALID study cohort consists of around 4000 people with DM2 of whom information on medical history, physical status, laboratory measurements, and medication have been prospectively collected. Patients who are being taken care of at the primary healthcare level in five European countries (Austria, Hungary, Netherlands, Poland, and Scotland) were recruited between 2011 and 2015 and followed for at least 5 years. Participants had to be aged 18–75 years and had an incident or prevalent DM2 defined as treatment with hypoglycaemic drugs or according to ADA guidelines at the time point of study entry. The presence of CKD at study entry was not an exclusion or inclusion criterion. For the full inclusion and exclusion criteria of patients, see Table 1. The aim of the study was to investigate regional differences in the course of diabetic nephropathy to determine the 5-year cumulative incidence of renal and cardiovascular outcomes and to identify predictive biomarkers for the eGFR trajectory in patients with DM2.

GCKD—German chronic kidney disease

The prospective cohort study GCKD consists of around 5000 people with CKD. Patients aged 18–74 years and suffering from CKD were included in case of an eGFR between 30 and 60 ml/min/1.73m^2 with or without urinary albumin excretion (CKD KDIGO Stage 3) or a better preserved eGFR in the presence of urinary albumin excretion >300 or protein excretion >500mg/day, and were followed up to 10 years. For the full inclusion and exclusion criteria of patients, see Table 1. Out of the 5000 recruited individuals, 1800 had diabetes and therefore represent the relevant patient subset for the prediction model in this work. The follow-up visits were set up in bi-yearly intervals. The main objective of the GCKD study cohort was to identify and validate risk factors for CKD, end-stage renal disease (ESRD), and cardiovascular disease (CVD) events.

DIACORE—the DIAbetes COhoRtE study

DIACORE is a prospective observational cohort study consisting of 6000 people with prevalent DM2 in Germany with at least 10 years of follow-up. The main objective of this study was the investigation of risk factors associated with the development and progression of diabetic complications through biosampling using high-throughput technologies for the collected biosamples (i.e., transcriptomics, proteomics, and metabolomics). To this end, patient information and blood samples were taken at baseline and at every bi-yearly follow-up visit for at least 10 years after study initiation.

Target population

The following inclusion criteria will be applied to the individuals from the PROVALID, GCKD, and DIACORE studies for the development and validation of the prediction model in this work:

- Caucasian ethnicity
- Presence of DM2 at baseline

Table 1 Overview of the inclusion and exclusion criteria of PROVALID, GCKD, and DIACORE

Study cohort	Study setting	
	Inclusion	**Exclusion**
PROVALID	▪ Aged 18–75 years ▪ Incident or prevalent DM2 with/without CKD	• Active malignancy requiring chemotherapy
GCKD	▪ Aged 18–74 years ▪ Caucasian ethnicity ▪ eGFR: 30–60 mL/min/1.73m^{2}with or without urinary albumin excretion ▪ eGFR > 60 mL/min/1.73m^2 - Albumin excretion of > 300mg/g creatinine - Proteinuria > 500mg/day	▪ Active malignancy in the 24 months prior to inclusion ▪ Renal or other transplantations ▪ NYHA IV heart failure
DIACORE	▪ Aged ≥ 18 years ▪ Caucasian ethnicity ▪ Prevalent DM2	▪ On chronic renal replacement therapy ▪ Active malignancy ▪ Pregnancy ▪ Further diseases and infections

- eGFR > 30 mL/min/1.73m^2 at baseline
- Aged 18–75 years at baseline
- At least three visits with recorded serum creatinine measurements (including baseline)
- At least 2 years of follow-up

Study outcome

The primary outcome of interest for the study is the annual decline in renal function, derived through the use of at least 3 follow-up measurements of eGFR. The eGFR values will be calculated by the equation of the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) using the patient's race, sex, age, and serum creatinine level [21].

Clinical predictor variables

The candidate variables for consideration in the prediction model were selected by medical experts. Variables of general clinical availability, ease in acquisition, and with clinical acceptance for inclusion in the model will be prioritized. Hence, the following variables will be included in the multivariable outcome model for the decline in renal function (see Table 2): age, sex, body mass index (BMI), smoking status (ever/never smoked), HbA_{1c}, urine albumin-to-creatinine ratio (UACR), presence of glucose-lowering medication, presence of lipid-lowering medication, presence of blood pressure-lowering medication, systolic and diastolic blood pressure, hemoglobin, and serum cholesterol. All predictors were measured at baseline, i.e., the first visit after the patient was included in either of the study cohorts. Due to the varying depth of information regarding the medication of an individual in the three cohorts, only drug indication classes will be included as a binary entry (y/n).

Sample size

Overall, 13 predictors will be included in the prediction model. Guidelines regarding the minimum required sample size for the development of a new multivariable prediction model have been recently proposed by Riley et al. [22]. Using the accompanying R package "pmsampsize" with an assumed average value for eGFR of 78.4mL/min/1.73m^2 and a standard deviation of 21.4 mL/min/1.73m^2 [23] in the population, an anticipated R^2 of the model of 0.3 [24] and a shrinkage factor of 0.9, the minimum sample size aimed at minimization of overfit and the precise estimation of the parameters of a prediction model with 13 predictors is 271. Further, since regularization is not taken into consideration in model development due to the abundance of data relative to the number of predictors, the computation with a shrinkage parameter of 0.99 results in a sample size of 3048 required subjects. However, a much larger R^2 is expected, which is why a minimum sample size of 1569 is required for a R^2 of 0.5 and 903 participants for 0.7, respectively. Nevertheless, the development cohort will comprise around 5800 subjects and thus will exceed the minimum required sample size by far.

Statistical analysis

Patient's baseline characteristics will be described for the study cohorts separately, using mean and standard deviation for continuous variables, or median and interquartile range in case of non-normality, and absolute and relative frequencies for categorical variables. Skewed variables will be transformed by the logarithm before the analysis. Data availability and the fraction of missingness will be assessed and reported for each variable. As the fraction of missing data is expected to be low, a

Table 2 Clinical baseline information of study participants within PROVALID, GCKD, and DIACORE

Variable type	Variable	Level of measurement	Unit
Demographics	Age	Continuous	Years
	Sex	Binary	Female-male
Physical status	Smoking	Binary	Never-ever
	BMI	Continuous	kg/m^2
Laboratory measurements	eGFR	Continuous	mL/min/1.73m^2
	Systolic blood pressure	Continuous	mmHg
	Diastolic blood pressure	Continuous	mmHg
	HbA_{1c}	Continuous	mmol/mol
	Serum cholesterol	Continuous	mg/dl
	Hemoglobin	Continuous	μmol/L
	UACR	Continuous	mg/g
Intake of medication	Blood pressure-lowering	Binary	Yes-no
	Glucose-lowering	Binary	Yes-no
	Lipid-lowering	Binary	Yes-no

complete case analysis will be carried out. Data screening through the computation of sample characteristics and visual inspections to identify the presence of missing values, skewed variable distributions, and outliers, as well as reporting of conducted screening activities will adhere to the conceptual framework for initial data analysis by Huebner et al. [25]. All analysis will be performed with the software R version 4.0.2.

Prediction model development

People adhering to the inclusion and exclusion criteria from either the PROVALID or GCKD study will be considered for model building. Mixed-effects modeling accounts for the dependencies between repeated measurements per person over time and the similarity among people belonging to the same study cohort, hence they are commonly employed to analyze longitudinal and clustered data. Here, we will consider hierarchically nested random effects for the country and people within a country, such that the heterogeneity of eGFR trajectories between the five countries within the development cohort can be modeled appropriately [26]. More specifically, we will use random effects for the intercept and the slope of the repeated eGFR measurements over time and an additional random intercept for the country. We will include the baseline eGFR value in the outcome vector to provide more stable estimates of variability. When applying the model, any available eGFR measurements of an individual for whom eGFR trajectory predictions should be obtained can be used to improve predictions. A similar approach has been conducted by Selles et al. [27] to develop a prediction model not only able to deal with multiple measurements per subject but also able to allow these repeated clinical measurements to contribute to the prediction of an unseen individual. The set of clinical variables outlined in Table 2 will constitute the fixed effects. Automated variable selection will not be conducted as the set of predictors was chosen by background knowledge. Furthermore, due to the large sample size available for estimation overfitting is expected to be minimal, such that further regularization of the predictor effects will be omitted. Clinically relevant pairwise interactions of the independent variables with time will be investigated and added if their inclusion improves the prediction accuracy of the model as determined by the Akaike information criterion (AIC). Any model misspecification, e.g., regarding functional forms of the variables in the model, will be assessed using a plot of the marginal residuals versus the individual variables. Variables with non-linear relationships with the outcome as exhibited by these residual plots will then be modeled using restricted cubic splines to improve the model fit [28]. In addition, the relevance of the inclusion of such non-linear terms for improving the model fit will be evaluated in terms of the AIC.

More formally, the model for the continuous eGFR values Y_{it} at time point t ($t = 0, \ldots, T$) for subject i ($i = 1, \ldots, n$) is defined by

$$\mathrm{Y_{it}} = \beta_0 + \mathrm{a_i} + \beta_1 X + \beta_2 \mathrm{t} + \beta_3 \mathrm{tX} + \mathrm{b_{0i}} + \mathrm{b_{1i}t} + \epsilon_{\mathrm{it}}, \tag{1}$$

with normally distributed residuals $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ with mean 0 and variance σ_ϵ^2, random country-specific intercepts assumed to be $a_i \sim N(0, \tau^2)$, the random coefficients $(b_{0i}, b_{1i}) \sim MN(0, G)$ (where MN denotes the bivariate normal distribution with an unspecified covariance matrix G, defined by the diagonal elements σ_0^2 and σ_1^2 and the off-diagonal σ_{01}) and the vector of regression coefficients for the clinical independent variables $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$. It should be noted that for simplicity of description, we here consider a single clinical predictor variable X, where β_1 and β_3 represent the regression coefficients corresponding to the clinical predictor variable and its interaction with time. Without loss of generality, more clinical predictors can be added, each with two regression coefficients for the baseline effects of the predictor and the corresponding interaction with time. When considering restricted cubic splines for a variable, this variable will be represented by several corresponding basis functions in (1).

Model (1) allows to prognosticate baseline and follow-up values of eGFR for a new person j given the value of the clinical predictor X, assuming that for that new person the random intercept and slope are equal to their expected values of 0. Further, we will assume the expected random country effect of $a_j = 0$ for the prediction of eGFR in a new individual. In order to incorporate a baseline eGFR measurement y_{j0} into the prediction, a best linear unbiased predictor (BLUP) of the individual's random effects can be obtained through the posterior distribution of the random effects. The random intercept for that new subject j is then estimated by

$$\hat{\mathrm{b}}_{0\mathrm{j}} = \frac{\sigma_0^2\left(\mathrm{y}_{\mathrm{j}0} - \hat{\mathrm{y}}_{\mathrm{j}0}\right)}{\sigma_0^2 + \sigma_\varepsilon^2}, \tag{2}$$

where $\hat{\mathrm{y}}_{\mathrm{j}0}$ denotes the estimated baseline value from the fixed effects, σ_ϵ^2 denotes the variance of the random error and σ_0^2 the variance of the random intercept in the covariance matrix of the random coefficients G. The estimated random slope $\hat{b}_{1j}$ for the new individual can be obtained by

$$\hat{b}_{1j} = \frac{\sigma_{01}\hat{b}_{0j}}{\sigma_0^2} \tag{3}$$

In actual predictions, estimates are plugged in for σ_o^2, σ_{01}, and σ_ε^2. The preliminary prediction $\hat{y}_{jt}$ for the new subject j at time t can then be updated by

$$\tilde{y}_{jt} = \hat{y}_{jt} + \hat{b}_{0j} + t\hat{b}_{1j} \tag{4}$$

The predicted eGFR slope for a new individual, combining fixed and random effects, is finally given by $d\tilde{y}_j \coloneqq d\tilde{y}_{jt}/dt = \hat{\beta}_2 + \hat{\beta}_3 X_j + \hat{b}_{1j}$ with $\tilde{V}_j$ expressing the variance of the predicted eGFR slope of individual j.

Given a correct model specification, the probability of progression can then be obtained by a normal approximation:

$$\Pr(Z = z_j) = 1 - \Phi\left(\frac{C_p - d\tilde{y}_j}{\sqrt{\tilde{V}_j}}\right) \tag{5}$$

where z_j, $j \in \{1, 2\}$ denotes the membership of belonging to one of the two progression groups: rapid and stable progression, and $\Phi(x)$ the standard normal distribution function at x. C_p is the corresponding subject-specific cutpoint for defining progression of kidney decline, which will be set to −3mL/min/1.73m^2/year at default or can be specified by the intended user in the web application (see Section "Web implementation") [14–16]. Now, one can compute the expected probability of belonging to either of the eGFR progression groups given a patient's baseline characteristics and the baseline eGFR.

For all statistical analyses, the freely available software R will be used. Mainly, the packages *nlme* and *JMbayes* will be considered for the implementation of the linear mixed model and the subject-specific prediction informed by the baseline eGFR measurement. The function *lme()* from the package *nlme* will be used to fit the linear mixed model and the function *IndvPred_lme()* from the package *JMbayes* to obtain the updated predictions and 95% prediction intervals for a new subject.

Model evaluation, performance, and external validation

An analysis of the conditional residuals will be conducted to check correct model specification, e.g., normality, linearity, and homoscedasticity of the error terms. Next, the importance of each independent variable will also be evaluated by the drop in the adjusted R^2, i.e., the loss in estimated explained variation when the respective predictor is removed from the model.

The performance of the prediction model will be evaluated by an internal-external validation procedure to obtain a reliable assessment of its generalizability of prediction to unseen data [29]. The unit for the non-random data splitting will be the six countries within the PROVALID and GCKD studies (GCKD: Germany; PROVALID: Austria, Hungary, Netherlands, Poland, and Scotland) such that for each iteration, one regional group with its study participants are withheld from model training and then used for testing.

The prediction performance will further be assessed through:

- Discrimination in terms of Kendall-tau-b concordance correlation coefficients at each time point, i.e., the concordance of the predicted eGFR values at time t obtained by Eq. (4) and the actual observed eGFR in each pair of patients
- Calibration in terms of the slope of the calibration curve by
  I. Regressing the predicted eGFR on the actual eGFR values at each time point
  II. Plotting the estimated mean risk for fast progression per decile against the observed ratio of fast progression within each risk decile
- Explained variation in terms of adjusted R^2, i.e., the proportion of explained variability in the outcome by the independent variables

The external validity of the model will be examined by applying the final prediction model to the DIACORE study cohort and evaluating the three performance measures outlined above.

Model presentation

Reporting of the model development and the final prediction model will adhere to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement, a checklist of items that ought to be reported when publishing a statistical prediction model [30]. The results of this analysis will be published in a peer-review journal. In particular, the final regression model will be presented using regression coefficients and 95% confidence intervals. Further, model reporting will include the covariance matrix of the random effects, the error variance, and the regression formula to allow independent application and validation of the model. Visualizations of results will be generated to improve model literacy for non-statistical readers and users of the web-based calculator.

Web implementation

The externally validated model will be implemented as an interactive prediction calculator, made available online as a web application to support clinical care and medical decision-making with regard to prevention and treatment of chronic kidney function decline in people with DM2. The interactive web application will be created with the help of the R development framework *shiny* to allow for individual input and the computation of patient-specific predictions given their baseline characteristics. The web application will offer two user interfaces to choose from:

(1) A patient interface with a web layout consisting of limited input controls (i.e., text fields or checkboxes to enter age, sex, eGFR, and UACR)
(2) A clinician interface with a layout featuring input controls for all predictors (i.e., includes laboratory measurements, medication) within the validated prediction model

Both interfaces will be based on the same prediction model formula, with predictors that are excluded from the web layout in the patient interface being fixed as the average value for the continuous variables and the value with the highest relative frequency for the categorical variables. As a result, the use of the model should not only be made available to clinical staff but also to patients themselves.

The output of the web application will contain the following components:

- The eGFR value estimated by the model in 1–5 years with 95% model-based prediction intervals
- The predicted eGFR decline per year with 95% model-based prediction interval
- The proportion of people with higher predicted eGFR loss
- The relative risk, i.e., the probability for a rapid eGFR decline for the predicted interval divided by the probability of a rapid eGFR decline for a person with equal age, sex, and baseline eGFR, but most favorable values of all other clinical baseline variables (with rapid decline specified by the intended user)
- A visual representation of the results:
- The visual communication of the results to the user will illustrate the estimated eGFR decline in the context of the distribution of predicted eGFR declines in the development cohort. In addition, a figure will be generated that shows the predicted trajectory of eGFR over the next 5 years with a 95% prediction interval.

A further option will be the possibility of risk assessment of rapid and stable eGFR decline. However, the choice of cutpoint selection for the categorization will be left to the user of the web-based implementation, in that the user can freely enter a suitable cutpoint for the eGFR classification into rapid and stable progression, whereby the risk of the new individual is estimated. By default, the eGFR cutpoint will be set to −3ml/min/1.73m^2.

Discussion

In this study, we have described the protocol for the planned development and validation of a new clinical prediction model for the decline of renal function in Caucasian people with DM2 to aid decision-making in clinical care. Our methodological approach will be based on a multivariable linear mixed model able to account for the dependence of clinical parameters per subject over time and for the similarity of individuals within the same country. Previous studies have employed risk prediction based on (multinomial) logistic regression models for which as a dependent variable the dichotomized eGFR slope was used, after estimation by least squares using the repeated measurements and then partitioning at arbitrary cutpoints. Each eGFR strata then covers different intervals on the eGFR slope spectrum and is supposedly indicative of varying severity of the progression of renal function loss. However, this approach suffers from several shortcomings. First, individual eGFR slopes can only be estimated very inaccurately, as only very few data points at non-equidistant time points are available for each patient [5]. Moreover, the categorization of the resulting eGFR slopes adds another layer of possible bias as it leads to subjects close on the continuous spectrum but on the opposite sides of the cutpoint being characterized as different [31]. By contrast, our approach avoids the issues associated with the imprecise estimation of slopes and inappropriate dichotomization prior to model development and takes the observed data fully into consideration when estimating the parameters of the model. In addition, we include the baseline measurements of eGFR into the outcome vector of an individual, as it is subject to the same measurement error as later eGFR measurements. In this way, these baseline measurements contribute to a more precise estimation of the error variance. Nevertheless, available eGFR measurements at the time of prediction (application of the model) can still be incorporated in our approach to optimize the predicted eGFR trajectory. In addition, the clustered data structure due to the two multicentre cohorts (GCKD and PROVALID) for model training will be taken into account by including an additional random intercept for the country in contrast to existing models. Therefore, the potential heterogeneity of model performance across the countries, which are distinct between GCKD and PROVALID, is prevented and hence, the generalizability to unseen individuals is improved.

However, this study will have a few limitations. First, it is to be expected that for some individuals not all creatinine measurements will be available, in particular, at later time points. Such random missingness is common in prospective longitudinal studies. Even in the case of informative missingness, by capturing much information about the patients by means of domain-expertise-selected clinical baseline variables, which are included in multivariable modeling, the mixed model is still able to provide unbiased predictions. Second, we will not include extensive biosampling data (e.g., metabolomics, proteomics, lipidomics) in the model. The inclusion of these biomarkers may improve the predictive accuracy of the model, but the scientific evidence for an added value in prediction is still scarce and the biomarkers have no validated test characteristics. We will rather prioritize clinical variables that are regularly collected and which are widely available to ensure the broad applicability of our model. Lastly, the lack of standardization of creatinine assays across cohorts inflicts variability of the clinical laboratory serum creatinine measurements and hence can induce potential bias in calculating eGFR.

The key strength of this study includes not only the refined methodology in the development of the prediction model, but also its large sample size due to the usage of two prospective observational cohort studies PROVALID and GCKD for model development. This ensures stable parameter estimation of the model. Another strength is the availability of an additional, independent prospective cohort study, DIACORE, for external validation after model development. To our knowledge, this is the first prediction model specifically developed for a central European population suffering from DM2 and covering a wide range of CKD stages.

Overall, we have outlined a robustly developed and validated clinical prediction model which will generalise to a wide range of patients with regard to initial CKD progression suffering from DM2. It will also be derived from a diverse and multinational population using two large studies and will be used to predict deterioration in renal function in DM2 patients to improve further eGFR development through early intervention in patients at high risk for the rapid progression of renal decline.

Conclusion

This will be a prediction model for eGFR loss in Caucasian subjects with DM2 that will use data from recent observational multinational studies. By adhering to transparent reporting procedures and current best practice for model development and validation, we will minimize the risk of bias when our prediction model is applied in the context of primary prevention of progression of CKD in people with DM2.

Acknowledgements
None.

Authors' contributions
MG was responsible for the study conceptualization and design and drafted the initial manuscript. GH was responsible for the study conceptualization and design and revised the manuscript critically for important intellectual content. AH, RO, MK, GM, CB, and KUE revised the manuscript critically for important intellectual content. HM has made substantial contributions to the acquisition of the data. The authors have approved the final submitted version and are accountable for all aspects of the work.

Funding
The research leading to these results has received support from the Innovative Medicines Initiative Undertaking under grant agreement no. 115974 BEAt-DKD.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
The PROVALID study protocol was approved in each participating country by the responsible local Institutional Review Board (IRB). Signing an informed consent was a prerequisite for study participation in all countries. The GCKD study protocol was approved by local institutional review boards at each participating academic institution, and the data protection concept was reviewed by the data protection officer of the State of Hessen. The DIACORE study protocol, the data protection strategy, and the study procedures have been approved by the Ethics Committees of participating institutions and are in accordance with the Declaration of Helsinki. Patients participate in DIACORE only after providing informed written consent.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria. [2]Division of Nephrology and Dialysis, Department of Internal Medicine III, Medical University of Vienna, Vienna, Austria. [3]Department of Nephrology and Hypertension, Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. [4]Department of Nephrology, University of Regensburg, University Hospital Regensburg, Regensburg, Germany. [5]Department of Nephrology and Medical Intensive Care, Charité Universitätsmedizin Berlin, Berlin, Germany. [6]Department of Internal Medicine IV (Nephrology and Hypertension), Medical University Innsbruck, Innsbruck, Austria.

Received: 26 July 2021 Accepted: 22 October 2021
Published online: 18 November 2021

References

1. Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. Lancet. 2021.
2. Jitraknatee J, Ruengorn C, Nochaiwong S. Prevalence and risk factors of chronic kidney disease among type 2 diabetes patients: a cross-sectional study in primary care practice. Sci Rep. 2020;10(1):1–10. https://doi.org/10.1038/s41598-020-63443-4.
3. Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: a review. Jama. 2019;322(13):1294–304. https://doi.org/10.1001/jama.2019.14745.
4. Pavkov ME, Collins AJ, Coresh J, Nelson RG. Kidney disease in diabetes. Diabetes in America 3rd Edition; 2018.
5. Kerschbaum J, Rudnicki M, Dzien A, Dzien-Bischinger C, Winner H, Heerspink HL, et al. Intra-individual variability of eGFR trajectories in early diabetic kidney disease and lack of performance of prognostic biomarkers. Sci Rep. 2020;10(1):1–7. https://doi.org/10.1038/s41598-020-76773-0.
6. Heerspink HJ, Stefánsson BV, Correa-Rotter R, Chertow GM, Greene T, Hou F-F, et al. Dapagliflozin in patients with chronic kidney disease. N Engl J Med. 2020;383(15):1436–46. https://doi.org/10.1056/NEJMoa2024816.

7. Ferreira JP, Kraus BJ, Zwiener I, Lauer S, Zinman B, Fitchett DH, et al. Cardio/kidney composite end points: a post hoc analysis of the EMPA-REG OUTCOME Trial. J Am Heart Assoc. 2021;10(7):e020053. https://doi.org/10.1161/JAHA.120.020053.
8. Bakris GL, Agarwal R, Anker SD, Pitt B, Ruilope LM, Rossing P, et al. Effect of finerenone on chronic kidney disease outcomes in type 2 diabetes. *N Engl J Med*. 2020;*383*(23):2219–29. https://doi.org/10.1056/NEJMoa2025845.
9. Dunkler D, Gao P, Lee SF, Heinze G, Clase CM, Tobe S, et al. Risk prediction for early CKD in type 2 diabetes. Clin J Am Soc Nephrol. 2015;10(8):1371–9. https://doi.org/10.2215/CJN.10321014.
10. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. Jama. 2011;305(15):1553–9. https://doi.org/10.1001/jama.2011.451.
11. Subasi E, Subasi MM, Hammer PL, Roboz J, Anbalagan V, Lipkowitz MS. A classification model to predict the rate of decline of kidney function. Frontiers in Medicine. 2017;4:97. https://doi.org/10.3389/fmed.2017.00097.
12. Vigil A, Condés E, Camacho R, Cobo G, Gallar P, Oliet A, et al. Predictors of a rapid decline of renal function in patients with chronic kidney disease referred to a nephrology outpatient clinic: a longitudinal study. Adv Nephrol. 2015;2015:1–8. https://doi.org/10.1155/2015/657624.
13. Heinze G, Dunkler D. Five myths about variable selection. Transpl Int. 2017;30(1):6–10. https://doi.org/10.1111/tri.12895.
14. Pena MJ, Heinzel A, Heinze G, Alkhalaf A, Bakker SJ, Nguyen TQ, et al. A panel of novel biomarkers representing different disease pathways improves prediction of renal function decline in type 2 diabetes. PLoS One. 2015;10(5):e0120995. https://doi.org/10.1371/journal.pone.0120995.
15. Eriksen B, Ingebretsen O. The progression of chronic kidney disease: a 10-year population-based study of the effects of gender and age. Kidney Int. 2006;69(2):375–82. https://doi.org/10.1038/sj.ki.5000058.
16. Shlipak MG, Katz R, Kestenbaum B, Fried LF, Newman AB, Siscovick DS, et al. Rate of kidney function decline in older adults: a comparison using creatinine and cystatin C. Am J Nephrol. 2009;30(3):171–8. https://doi.org/10.1159/000212381.
17. Rifkin DE, Shlipak MG, Katz R, Fried LF, Siscovick D, Chonchol M, et al. Rapid kidney function decline and mortality risk in older adults. Arch Intern Med. 2008;168(20):2212–8. https://doi.org/10.1001/archinte.168.20.2212.
18. Eder S, Leierer J, Kerschbaum J, Rosivall L, Wiecek A, de Zeeuw D, et al. A prospective cohort study in patients with type 2 diabetes mellitus for validation of biomarkers (PROVALID)–study design and baseline characteristics. Kidney Blood Press Res. 2018;43(1):181–90. https://doi.org/10.1159/000487500.
19. Eckardt K-U, Bärthlein B, Baid-Agrawal S, Beck A, Busch M, Eitner F, et al. The German chronic kidney disease (GCKD) study: design and methods. Nephrol Dial Transplant. 2012;27(4):1454–60. https://doi.org/10.1093/ndt/gfr456.
20. Dörhöfer L, Lammert A, Krane V, Gorski M, Banas B, Wanner C, et al. Study design of DIACORE (DIAbetes COhoRtE)–a cohort study of patients with diabetes mellitus type 2. BMC Med Genet. 2013;14(1):25. https://doi.org/10.1186/1471-2350-14-25.
21. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF III, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604–12. https://doi.org/10.7326/0003-4819-150-9-200905050-00006.
22. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: part I–continuous outcomes. Stat Med. 2019;38(7):1262–75. https://doi.org/10.1002/sim.7993.
23. Bramlage P, Lanzinger S, Hess E, Fahrner S, Heyer CH, Friebe M, et al. Renal function deterioration in adult patients with type-2 diabetes. BMC Nephrol. 2020;21(1):1–10. https://doi.org/10.1186/s12882-020-01952-0.
24. Heinzel A, Kammer M, Mayer G, Reindl-Schwaighofer R, Hu K, Perco P, et al. Validation of plasma biomarker candidates for the prediction of eGFR decline in patients with type 2 diabetes. Diabetes Care. 2018;41(9):1947–54. https://doi.org/10.2337/dc18-0532.
25. Huebner M, le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. Observational Studies. 2018;4(1):171–92. https://doi.org/10.1353/obs.2018.0014.
26. Falconieri N, Van Calster B, Timmerman D, Wynants L. Developing risk models for multicenter data using standard logistic regression produced suboptimal predictions: a simulation study. Biom J. 2020;62(4):932–44. https://doi.org/10.1002/bimj.201900075.
27. Selles RW, Andrinopoulou E-R, Nijland RH, Van Der Vliet R, Slaman J, van Wegen EE, et al. Computerised patient-specific prediction of the recovery profile of upper limb capacity within stroke services: the next step. J Neurol Neurosurg Psychiatry. 2021;92(6):574–81. https://doi.org/10.1136/jnnp-2020-324637.
28. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagnostic and Prognostic Research. 2020;4(1):1–18. https://doi.org/10.1186/s41512-020-00074-3.
29. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.
30. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation. 2015;131(2):211–9. https://doi.org/10.1161/CIRCULATIONAHA.114.014508.
31. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080. https://doi.org/10.1136/bmj.332.7549.1080.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.