Study protocol
Dynamic risk predictions in adults with moderate-severe chronic kidney disease

Investigators

Alberta
Mr. Emir Sevinc (lead), University of Calgary, Alberta, Canada
Dr. Tyrone Harrison (clinical epidemiology co-supervisor), University of Calgary, Alberta, Canada
Dr. Pietro Ravani (clinical epidemiology supervisor), University of Calgary, Alberta, Canada

Aberdeen
Dr. Simon Sawhney, University of Aberdeen, Scotland, UK
Dr. Andrew Mclean, University of Aberdeen, Scotland, UK

Denmark
Dr. Christian Fynbo Christiansen, Aarhus University Hospital, Denmark
Dr. Simon Kok Jensen, Aarhus University Hospital, Denmark
Dr. Uffe Heide-Jørgensen, Aarhus University Hospital, Denmark
Dr. Thomas Alexander Gerds (statistical supervisor), University of Copenhagen, Denmark

Version
Version1_20230719

Motivation of this work

Risk prediction tools are used to help people who are facing difficult treatment decisions better understand their chances. People who are diagnosed with chronic kidney disease (CKD) want to know how likely their kidneys are to fail at some point in the future, to decide whether and how to prepare for the management of kidney failure. As in any chronic condition, also in people with CKD mortality simultaneous assessment of mortality risk is key to inform treatment decisions. Given the high risk of death in the CKD population, accurate prediction of both risks of kidney failure and death is necessary to facilitate tailored clinical decision-making and preparations beyond those solely related to the management of kidney failure.

We recently created KDpredict (http://kdpredict.com), a tool for predicting the 1-to-5-year risks of kidney failure and death in adults with moderate-to-severe chronic kidney disease (G3b and G4 CKD, G3bG4-CKD) from disease onset (first documentation of G3bG4-CKD). KDpredict includes 4 cause-specific Cox (CSC) models for the prediction of kidney failure and 4 standard Cox (COX) models for death. For each outcome, 2 models include 4 predictors (age, sex, albuminuria and eGFR calculated the CKD-EPI 2009 equation or age, sex, albuminuria and eGFR calculated the CKD-EPI 2021 equation) and 2 models include 6 predictors (the same as above with the addition of cardiovascular disease and diabetes). These 8 models were selected using a discrete super-learner algorithm fitted with 8 libraries of base learners. Each library included CSC or COX models with different specifications of the predictors and random forest algorithms for competing risks or standard survival data. KDpredict was trained in Alberta and tested in Scotland and Denmark. For the outcome of kidney failure, KDpredict was compared to the existing benchmark model kidney failure risk equation (KFRE). For the outcome of death, no comparison was possible because no survival model existed at that time.

KDpredict was created using a discrete super-learner, i.e., for each library the super-learner algorithm selected one winner, the model with the smallest prediction error (Brier score). The super-learner library included many different Cox regression models and random survival forests built on 4 to 6 predictors measured at cohort entry. Additional work may clarify if we can improve the performance and usability of KDpredict by considering: (1) additional predictors, including indicators of health resource use or kidney function changes before cohort entry; (2) a super-learner ensemble[1] instead of a discrete super-learner approach, that may include parametric regression models, classification trees and deep learning algorithms; and (3) models for updated (dynamic) individual predictions over time. We will build a series of learning ensembles that predict the 1-to-5-year risks of kidney failure and all-cause mortality in people who have moderate-to-severe CKD from a common time origin (first diagnosis) and at different landmark points in survivors (e.g., years 1-to-2 from cohort entry).

Methods for super-learning have been developed for continuous and categorical outcomes but are in their infancy for time to event outcomes and do not exist for competing risks. Also, there is no benchmark tool for dynamic risk prediction in people with CKD. For these reasons, we will first consider a study based on super-learning ensemble for standard time to event data before we consider the competing risks setting.

In Study 1 we will create a prediction ensemble and compare its performance to the discrete super-learner for the outcome of survival (time to death). This will be a static prediction tool for 1-to-5-year risk predictions from cohort entry (first encounter). We will consider the same predictors of the

original KDpredict as well as hospitalization data, emergency visits and estimated glomerular filtration rate (eGFR) changes in a window of 1-to-3 years before cohort entry (prediction time origin).
In Study 2 we will create a prediction tool that can be used at time zero (disease onset) and in years 1, 2 and 3 from time zero for the prediction of death. Predictors will be updated at each landmark time.
In Study 3 and Study 4 we will implement the same approach to predict the outcome of kidney failure accounting for the competing risk of death.

Prediction framework

Target population

CKD is defined as the presence of abnormal levels of albuminuria or estimated glomerular filtration rate (eGFR) below 60 mL/min/1.73 m^2 for more than 90 days. We will include adults with G3bG4-CKD (eGFR 15-44 mL/min/1.73 m^2): G3b-CKD (eGFR 30-44 mL/min/1.73 m^2) and G4-CKD (eGFR 15-29 mL/min/1.73 m^2) only. We will exclude G3a-CKD (eGFR 45-59 mL/min/1.73 m^2) for the following reasons. First, people with G3a-CKD (the largest fraction of people with CKD, >60%) have very low risk of kidney failure (<1% at 5 years). For these people risk prediction at the 1-5-year time horizons is of unclear clinical interest. Instead, most of them need diagnosis and evidence-based treatments. Second, most people meeting criteria for G3a-CKD may simply have 'kidney ageing' rather than CKD, as they do not have albuminuria and are older than 65 years. Third, people with G3a are difficult to capture when implementing a sustained definition of CKD as serum creatinine (eGFR is calculated from age, sex and serum creatinine) tends to be measured less often with less severe CKD.[2] This may introduce systematic differences by CKD stage in the definition of time zero for survival analyses.

We will capture G3bG4-CKD using an algorithm that identifies people with outpatient eGFR measures between 15 and 44 ml/min/1.73 m^2 sustained for 90 days, as per practice guideline's recommendations. We will exclude stage 5 CKD (eGFR <15 ml/min/1.73 m^2) and people with previous history of kidney failure.

Intended use

The static predictor tool will be used once at disease onset. The dynamic prediction tool will be used at baseline as well as at years 1 and 2 from baseline.

Outcome

The outcome will be death from all causes for studies 1 and 2; and kidney failure for studies 3 and 4.

Prediction parameter

Individualized risk prediction.

Prediction origin

Time zero (cohort entry or G3bG4-CKD onset) for the static predictor tool and time zero, 1, 2 and 3 years from onset for the dynamic predictor.

Prediction horizon

We will consider years 1, 2, 3, 4 and 5 for the static predictor. Only years 1 and 2 for the dynamic predictor (prediction updates).

Existing tools have several methodological limitations

Grams et al.[3] meta-analysed HRs (or sub-HR) from models built in 29 cohorts of the CKD consortium and derived the pooled underlying hazard (sub-hazards, which they incorrectly called 'absolute risk') assuming a Weibull distribution.[3] The model is for people with G4-CKD only.
Limitations of the KFRE have been extensively discussed.[4-6]

Objective

To train and test different algorithms to predict the risk of death in Alberta and externally evaluate the best risk prediction tool in Ontario, Denmark and Scotland.

Commented [ES1]: Thomas: Unclear objective. It shouldnt say how we do it, it should say what we wish to achieve. How do we define "best"? OR best in Alberta, or separately in Ontario, DK and SC? IT could very well be that the best fitting model here could be poor for the others. Might as well create a separate model in each location. Super learner comparison (Discrete vs Ensemble) should first be tested with simulated data. How would we generate simulated data such that the meta learner outperforms the discrete learner?
1 - all models we put in are wrong (none of them generate our data)
2 - the sample size is too small to estimate the data generating mechanism well

Commented [ES2]: One way to create scenario 1: introduce an unobserved confounder variable, so that all of the models are wrong naturally. The discrete will be the best if the data generating model is a pat of the discrete library. Since we're synthesizing data, we will know the data generating model. If we use only the real data, then maybe all models in the library are wrong (we wont know). And then, the Brier score can not decide between two different wrong models. In order to choose between two different wrong models, we need a clinical implication of the decisions. Cost vs Benefit. Medical decisions could be starting the treatment or not, etc. There are costs (side effects, monetary etc) to treatments, and benefits (lower mortality etc).

Approach

We will include stable, incident G3bG4-CKD to reflect the characteristics of the prediction tool users. We will focus on a pre-specified target of predictions (individual risk) at time horizons seen from a common time origin. The algorithm to identify CKD cases using eGFR will be designed in Alberta and implemented in other sites. For albuminuria, we will use different methods, including albumin-to-creatinine ratio (ACR) and methods from which ACR can be calculated (protein-to-creatinine ratio, PCR, and urine dipstick). Urine dipstick is part of usual care and workflow for the information management system in Alberta and Ontario, but not in Denmark and Scotland. However, we will include people who had only urine dipstick measures of proteinuria in the training cohort (Alberta). This approach will (1) minimize the number of people excluded from the study and (2) enhance applicability of the prediction tool to users who have access to urine dipstick test results. Since inclusion of ACR derived from urine dipstick may reduce model performance, we are aware that further studies will be needed to test the model performance.

Methods

Date range for cohort entry

2008, April 1 to 2019, March 31 (more recent data may become available in the next year).

Study end date

2020, March 31 (more recent data may become available in the next year).

General approach

We will screen all eGFR data in the lab repository of each jurisdiction. Criteria for cohort entry (incident stage G3bG4-CKD) and development of chronic kidney failure (kidney outcome), will be an eGFR reduction sustained for >90 days (recommended chronicity criterion) below the thresholds of 45 and 10 mL/min/1.73 m^2, respectively.

In studies 1 and 2, patients will not be censored if they meet criteria for kidney failure during follow-up from time origin or from any landmark time. For the dynamic tool, however, survivors who develop kidney failure will be excluded from the new landmark analysis (a separate mortality model will be developed for people with kidney failure), as they will no longer meet entry or re-entry criteria. In studies 3 and 4, the outcome will be time to kidney failure, accounting for the competing risk of death.

Given the low prevalence of people in each jurisdiction who self-identified as being Black individuals, we will calculate eGFR using the CKD-EPI-2009 equation to identify participants and capture outcomes (which is more accurate than the 2021 equation in non-Blacks), excluding the race coefficient, with serum creatinine values standardized to isotope dilution mass spectrometry-traceable methods. We will use only outpatient eGFR measurements to reflect the characteristics of the target population and minimize the inclusion of people with episodes of acute kidney injury or unstable clinical conditions. We will use the mean value of eGFR when there are multiple measurements on the same day. We will also calculate the baseline eGFR from the index serum creatinine using the CKD-EPI-2021 equation to obtain a version of each prediction algorithm with eGFR calculated without race.

Cohort

To identify people with sustained reduction of eGFR meeting criteria for cohort entry (moderate-severe, i.e. G3b-CKD and G4-CKD), we will screen each individual's series of ≥2 consecutive eGFR tests where the first and last eGFR measurements are separated by >90 days, the first and all possible intervening measurements within 90 days are <45 mL/min/1.73 m^2 and the last eGFR is 15-44 mL/min/1.73 m^2. We will select the earliest series (qualifying period) where all eGFR measurements will meet the eGFR requirement for cohort entry (Table 1). The date of the last eGFR in the qualifying period will define the index date (cohort entry date or prediction time origin). We will exclude people with previous history of maintenance dialysis or kidney transplant or sustained eGFR <15 mL/min/1.73 m^2 for more than 90 days (stage 5 CKD) on or before the index date, because they are no longer at risk of chronic, end-stage kidney failure of their native kidneys. Albuminuria is a required input in the current kidney failure risk prediction tool. We will include people with records of urine albumin-to-creatinine ratio, urine protein-to-creatinine ratio or urine dipstick in the 3 years preceding cohort entry. We will exclude individuals without proteinuria measurements within 3 years before the index date, because missing information on proteinuria is rarely random (people without proteinuria information tend to be older and have higher risk of death and lower risk of kidney failure) and thus data imputation may be problematic.

Outcomes

The outcome will be death from all causes for studies 1 and 2; and kidney failure for studies 3 and 4 (Table 2).

Predictors

Baseline age, sex, index eGFR (calculated using the 2009 formula for the 2009 super-learner libraries and the 2021 formula for the 2021 super-learner libraries), albuminuria, diabetes, and cardiovascular disease (one or 4 variables including congestive heart failure, myocardial infarction, peripheral vascular disease, stroke or transient ischemic attack). Cancer (except epithelial skin cancer) and chronic pulmonary diseases will be used only for cohort description (Table 3). We will use the most recent outpatient proteinuria values within 3 years before study entry, with the following types of measurement in descending order of preference: albumin-to-creatinine ratio, protein-to-creatinine ratio, and urine dipstick. We will use the work by Sumida et al to convert PCR to ACR (crude formula).16 If there are ≥2 same type measurements on the same day, we will use the median value of ACR or PCR measurements. For urine dipstick protein measurements, we will apply the floor function of median category that returns the largest integer less than or equal to a given value. We will use validated algorithms to identify the comorbidities listed above. Predictor values will be updated at each landmark time in landmark analysis. We will also consider predictors related to health care resource use and eGFR changes before cohort entry. The simplest was to summarize this information is to use a binary variable, admission in the year before cohort entry (yes/no). This can be used without access to electronic medical records. For automatic use of recorded health data we will consider the days spent in hospital and the number of emergency visits within 1-to-3 years of cohort entry and an eGFR below 15 mL/min/1.73 m^2 during the qualifying period.

Analysis plan

We will use standard descriptive methods to summarize baseline characteristics by jurisdiction, including age, sex, eGFR (calculated from 2009 or 2021 equation), qualifying period (days), number of eGFR tests during the qualifying period, albumin-to-creatinine ratio (mg/g and dummies generated from ACR, A1, A2, and A3), diabetes, cardiovascular disease (myocardial infarction, heart failure, stroke/TIA, or peripheral vascular disease), chronic pulmonary disease (chronic lung disease or asthma), and cancer (except epithelial skin cancer). We will summarize mortality rates and follow-up times in a Table or text depending on the journal requirements. We will use the reverse Kaplan-Meier estimator for the censoring distribution to summarize the follow-up time.

Super-learner design

KDpredict was created using a discrete super-learner approach. We will consider a super-learner ensemble for this project.[1] Super-learning is a cross-validation procedure that selects the best performing algorithm among many alternatives (base learners) based on its ability to minimize prediction error. EXPLAIN ENSEMBLE HERE, GENERAL CONCEPT.

We will have two separate analysis phases in Alberta, one using synthetic data without access to the observed outcome and one with access to the observed outcome (Figure). We will use the synthetic data for unsupervised learning (e.g., for random forest tuning, also known as hyperparameter optimization in an outcome blinded fashion). The super-learner will be trained using the entire (supervised learning). We will study of the max prediction times using bootstrapping of survival times in the last landmark cohort of survivors.

In unsupervised learning, we will use clinical knowledge and a synthetic sample with the same probability distribution of the combinations of the predictor variables as the entire Alberta cohort, to design the base learners of each super-learner library (Tables 4-5). The outcome of the synthetic data will be altered with random numbers (label-blinded learning). Base learners will include Cox proportional hazard models, with 4 or 6 variables and different settings (with or without interaction terms or spline transformations of continuous variables). We will also include parametric Weibull regression. Explain settings. We will consider the same 4 or 6 variables to grow random survival forests the outcome of death. We will use the synthetic dataset to tune these random forests for node size and number of variables for random bootstrap sampling at each split to minimize out-of-bag error using 100-500 trees per ensemble. The flexibility of a random survival forest allows the relaxation of the assumptions of proportional hazards, linearity of continuous predictor variables, and no interactions between predictor variables, and thus provides indirect evidence on the goodness of fit of a rival semi-parametric model in head-to-head comparison. We will consider the following additional machine learning algorithms: Survival Neural Networks (Survnet), and Extreme Gradient Boosting (XGBoost).

In supervised learning, we will use the entire Alberta cohort to identify the best performing base learner by fitting a discrete super-learner with the pre-specified base learner library. After that, the best performing **discrete learners** will be compared with an **ensemble learner**. An ensemble learner is a method that uses predictions from multiple base learners to obtain improved predictions. Risk predictions are produced from each base learner, after which ensemble predictions are produced as a function of the base predictions. This ensemble super-learner combines the best-performing learners into a new, more accurate, risk prediction. We plan to use all the previously mentioned methods, i.e., Cox and Weibull regression models, Random Forests, Neural Networks and XGBoosting as base learners for the ensemble approach.

We will use internal cross-validation based on 500 bootstrap sets each obtained by random subsampling 63.2% of the full internal cohort for learning and 36.8% to calculate the prediction performance for both the base learners and the ensemble method. We will use the leave-one-out bootstrap for averaging the results across multiple splits. The winner (which could be an ensemble or a discrete learner) will have the lowest time-dependent Brier score. From each version of the super-learner (4- and 6-variableand with eGFR calculated using the CKD-EPI-2009 and CKD-EPI-2021 formulas), we will obtain the outcome-specific winner with the lowest mean Brier score over all time horizons. We will repeat this cross-validation scheme for three formulations of the super-learner: (1) 4-variable super-learner: with 4 predictors only (sex, age, eGFR, ACR); and (2) 6-variable super-learner: the same 4 variables plus diabetes and cardiovascular disease summarized as binary variable. We will obtain a super-learner including the eGFR calculated using the 2021 equation instead of the 2009 equation. We will consider use of health care resources and eGFR changes before cohort entry as additional predictors.

Established methods exist for ensemble super-learner with generalized linear outcomes. The method to estimate the coefficients for the ensemble super-learner (AKA ensemble weights) have been recently developed for survival data (e.g., survSuperLearner and SuperLearner-survival packages in R) but they are still in their infancy. We will test the discrete super-learner against an ensemble that will include a Cox model, a parametric Weibull model, a random forest, an XGboost tree ensemble, and a neural network for survival data using new algorithms that we are currently developing. To create this ensemble super-learner, we will first obtain a discrete super-learner for each model and algorithm (i.e., the best Cox, the best tree ensemble, etc). Then, we will use a super-learner algorithm to obtain the final ensemble.

Dynamic vs. static prediction tools. Given the need to update predictions in survivors over time, we will consider updated time origins (landmark times) at years 1-3 from baseline, with updated covariates (including eGFR or kidney failure). In landmark analysis we will include only people who are event-free at each landmark time.

Evaluation of the prediction tool

Internal testing. We will use the same cross-validation scheme to summarize the performance of different formulations of the super-learner in the learning cohort (for 1-5-year predictions). We will not present these internal testing results in the paper, because the focus will be on external transportability. Instead, we will consider cross-validation to assess the performance of retrained models where necessary (see below “Other analysis plans”).

Spatial transportability. We will use the full sets of external cohorts to investigate to what extent the super-learner trained in Alberta could be exported to Denmark and Scotland.

Temporal testing. We will split the Alberta cohort in a training and testing set by calendar data (NEED TO SAY). The SL will be retrained in the training set and tested using the unseen data.

Comparisons

We will compare the performance of the SL ensemble with the discrete SL previously created (original KDpredict).

Performance measures

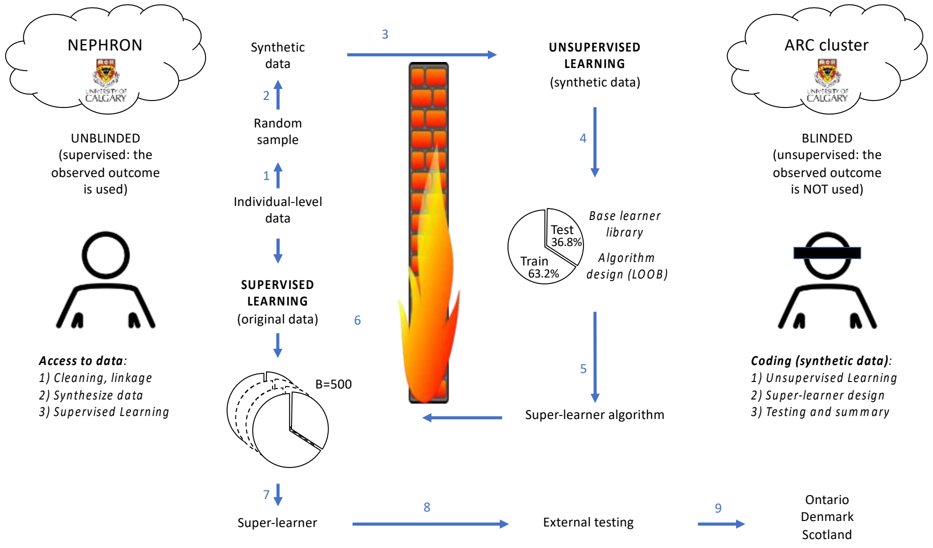
1) Prediction agreement. We will use scatterplots to assess the agreement between individualized risk predictions for all comparisons (listed above).
2) We will use calibration in the small (histogram-type plots) and time-dependent AUC, Brier score, and index of prediction accuracy for model comparison.
a. For calibration, we will categorize the predicted risks from each model for all individuals in each testing dataset into a pre-specified number of equally large groups (predicted risk deciles) and compare for each group the mean predicted risk to the mean estimated actual risk. A model is well calibrated if the two means have the same height in all groups.
b. We will use the inverse probability of censoring weighted estimates of the area under the time-dependent receiver operating characteristic curve (AUC; the higher the better) to estimate model discrimination. Being a ranking statistic, the AUC cannot stand alone to assess models with respect to predictive accuracy.
c. We will use the Brier score (prediction error, a measure of both calibration and discrimination; the lower the better) and index of prediction accuracy (IPA, a measure of average prediction performance derived from the Brier score; the higher the better). Note that among these predictive performance measures (scores), we will favour the use of the Brier score, which is the only strictly proper scoring rule.
d. Since two rival models may predict very different individual risks yet differ only slightly in their scores (the Brier score reflects both calibration and discrimination), we will consider clinically relevant changes (>10% difference) in individualized predicted risks, calibration and the numeric value of the Brier score more meaningful than the size of the differences in performance scores.

Other analysis plans. We will consider refitting the super-learner also in external testing sites (Ontario, Denmark and Scotland) in case of suboptimal external performance. Testing the SL using unseen data will require temporal testing as described for the Alberta cohort.

Software and packages

We will use R (version >4.1.0) for all analyses and the main packages riskRegression and randomForestSRC. ADD ANY.

Figure: Super-learning design



Legend: Individual-level data will be stored within the Alberta Kidney Disease Network (AKDN) server and kept inaccessible to researchers to minimize the risk of contamination and over-fitting. Step 1: Researchers will give instructions to AKDN analysts on how to prepare the data and obtain a random sample from the original dataset. Steps 2-5: A synthetic dataset will be created and sent to the Advanced Research Computing (ARC) cluster for unsupervised learning (label-blinded learning). The synthetic sample will have the same probability distribution of the combinations of the predictor variables as the Alberta neutral cohort and time to event altered by random numbers (outcome blinding). Unsupervised learning will consist in the creation of a library of base learners (feature analysis and model design) and design of super-learner algorithms. The super-learner will use (1) cross-validation based on 500 bootstrap sets each obtained by random subsampling 63.2% of the full cohort for learning (in-bag) and 36.8% to calculate the prediction performance (out-of-bag) and (2) the leave-one-out bootstrap (LOOB) for averaging the performance results across multiple splits. Three sets of predictors (4-, 6- and 11-variables) will be considered, each with 2 versions, one with eGFR calculated with the 2009 formula and one with the 2021 formula. For each of these sets of predictor variables (3X2), the algorithm will select the base learner with the lowest mean of the 5 cross-validated time-dependent Brier scores (each separately obtained at 1-, 2-, 3-, 4-, and 5-year time horizons). Steps 6-7: The R scripts will be sent back to the AKDN analysts for supervised learning (to implement the procedure using the original data). Steps 8-9: Sensitive data will be removed from the R objects (the 6 models) that will be sent to the external sites for testing.

Table 1. Sustained eGFR methods for cohort formation and kidney outcome ascertainment

	Qualifying period (>90 days)		
Cohort	*First eGFR*	*Intervening eGFR*	*Last eGFR (index date)*
G3b-G4 CKD	<45	<45	≥15 & <45
Outcome	*First eGFR*	*Intervening eGFR*	*Last eGFR (outcome date)*
KF_{eGFR}	<10	<10	<10

Legend: eGFR, estimated glomerular filtration rate (mL/min/1.73 m^2) using the 2009 formula.

- G3b-G4 CKD refers to GFR categories of CKD. G3b CKD: CKD with moderately to severely decreased GFR; G4 CKD: CKD with severely decreased GFR. For simplicity, we call them moderate and severe CKD, respectively.
- Qualifying period for study entry: the earliest period for at least two consecutive eGFR measurements <45 mL/min/1.73 m^2 for >90 days.
- Intervening eGFR: any eGFR (from 0 to n) between the first and the last eGFR of the qualifying period.
- Index date (prediction time origin): date of the last eGFR of the qualifying period for study entry.
- Index eGFR (baseline eGFR): the last eGFR of the qualifying period for study entry.
- Kidney failure (KF) state (time-varying covariate for landmark analysis) defined as the earliest of initiation of kidney transplant, maintenance dialysis or sustained eGFR <10 mL/min/1.73 m^2 for >90 days.

Table 2. Codes for identifying kidney transplantation using administrative data (Alberta)

1) Physician claims: Canadian Classification of Diagnostic, Therapeutic, and Surgical Procedures codes

Codes	Code description
67.5	Transplant of kidney
67.59	Other kidney transplantation
67.59A	Renal transplantation (homo, hetero, auto)

2) Hospitalizations

Codes	Code description
Canadian Classification of Health Intervention codes	
1.PC.85.^^	Transplant, kidney
1.PC.85.LA-XX-J	Using living donor (allogenic or syngeneic) kidney
1.PC.85.LA-XX-K	Using deceased donor kidney
1.OK.85.XU-XX-K	Transplant, pancreas with duodenum and kidney with exocrine drainage via bladder [e.g. donor duodenum is grafted to bladder: duodenocystostomy]
1.OK.85.XV-XX-K	Transplant, pancreas with duodenum and kidney with exocrine drainage via intestine with homograft [e.g. donor duodenum is grafted to bowel]
ICD-9-CM procedure codes	
55.69	Other kidney transplantation

Table 3. Codes for identifying comorbidities using administrative data

Comorbidities		Algorithm	ICD-9 CM	ICD-10	Additional sources (site)
Diabetes		1 hospitalization or 2 claims in 2 years or less	250	E10-E14	Medications: Denmark, ATC codes A10A and A10B from prescription registry Grampian medication BNF codes 060101, 060102 from previous year
Cardiovascular disease	Myocardial infarction	1 most responsible hospitalization	410	I21-I22	-
	Chronic heart failure	1 hospitalization or 2 claims in 2 years or less	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4–425.9, 428	I09.9, I25.5, I42.0, I42.5–I42.9, I43, I50	-
	Stroke or transient ischemic attack	1 most responsible or post-admittance hospitalization or 1 claim or 1 most responsible ED ACCS	362.3, 430, 431, 433.×1, 434.×1, 435, 436	G45.0-G45.3, G45.8-G45.9, H34.1, I60, I61, I63, I64	-
	Peripheral vascular disease	1 hospitalization or 1 claim or 1 ACCS	440.2	I70.2	-
Chronic pulmonary disease	Chronic lung disease	1 hospitalization or 2 claims in 2 years or less	416.8, 416.9, 490–492, 494–505, 506.4, 508.1, 508.8	I27.8, I27.9, J40–J44, J46-J47, J60–J67, J68.4, J70.1, J70.3	-
	Asthma	1 hospitalization or 3 ACCS in 2 years or less	493	J45	Medications: Northern Denmark, ATC code R03 from prescription registry Grampian BNF codes 030100, 030201, 030202 from previous year
Cancer		1 hospitalization or 1 claim, look back for 5 years from index date for diagnoses	140-209, except 173	C00-C96, except C44	-

Table 4. Library of base learners

*Base learners	#N of variables	Strata (stage)	^Splines	Interactions	N trees	Features tried at each node	Split N	Node size
COX	4, 6	Yes or no	From none to 3 (ACR, age, eGFR)	Consider first order only	-	-	-	-
RSF	4, 6	-	-	-	100-500	Ceiling(SQRT(X) +/-1	10-15	10-15

Legend:
*Base learners: COX: Cox regression; RSF: random survival forest for survival data
We are currently developing the strategy for XGBoost and neural network algorithms

#N of variables (input features):
4-variable super-learner: age, eGFR, log-ACR and sex
6-variable super-learner: age, eGFR, log-ACR, sex, diabetes, any cardiovascular disease
11-variable super-learner: age, eGFR, log-ACR, sex, diabetes, heart failure, myocardial infarction, peripheral vascular disease, stroke or transient ischemic attack, cancer, chronic pulmonary disease

The base-learner library will include many base learners characterized by different settings for the COX (different combinations of spline number or interactions). For the random forest, we will tune the hyperparameter ‘features tried at each node’ as a function of the total number of features. Other hyperparameters we will consider varying include the N of trees, N of splits and node size. One version of the super-learner will include eGFR calculated with the CKD-EPI 2009 equation, and one the CKD-EPI 2021 equation.

^Splines: restricted cubic splines, with estimated or pre-specified knots:
Age: 65, 75, and 85 years
eGFR: 25, 35 and 40 mL/min/1.73 m^2
log-ACR: log(30), log(300) and log(600), where ACR (mg/g) is albumin-to-creatinine ratio

Table 5. Names of variables

Name	Units/levels	Description
death#	Binary: 0,1	All cause mortality
year#	Continuous, in years	Time to death from baseline
male	0,1	Sex
age	years	Age at baseline
gfr^	ml/min/1.73 m2	Baseline eGFR
lACRc	Log(mg/g)*	Baseline log-ACR (ACR or calculated from PCR mg/g* using crude model)
dm	0,1	Baseline diabetes
cvd	0,1	Baseline cardiovascular disease
ami	0,1	Baseline myocardial infarction
chf	0,1	Baseline congestive heart failure
pvd	0,1	Baseline peripheral vascular disease
stroke	0,1	Baseline stroke
cpd	0,1	Baseline chronic pulmonary disease
cancer	0,1	Baseline cancer

# names:
year09, death09 = eGFR input calculated using the 2009 formula
year21, death21 = eGFR input calculated using the 2009 formula

^ names:
gfr09 = index eGFR calculated with the 2009 formula
gfr21 = index eGFR calculated with the 2021 formula

*Conversion: 1 mg/mmol = 0.113 mg/g.

Pre-entry information on health care resource use and eGFR changes to add.

References

1. Laan MJvd, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology 2007;6.
2. Liu P, Quinn RR, Lam NN, et al. Progression and Regression of Chronic Kidney Disease by Age Among Adults in a Population-Based Cohort in Alberta, Canada. JAMA Netw Open 2021;4:e2112828.
3. Grams ME, Sang Y, Ballew SH, et al. Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate. Kidney international 2018;93:1442-51.
4. Al-Wahsh H, Tangri N, Quinn R, et al. Accounting for the Competing Risk of Death to Predict Kidney Failure in Adults with Stage 4 Chronic Kidney Disease. JAMA Network Open 2021;4:1-13.
5. Hundemer GL, Tangri N, Sood MM, et al. The Effect of Age on Performance of the Kidney Failure Risk Equation in Advanced CKD. Kidney Int Rep 2021;6:2993-3001.
6. Ravani P, Fiocco M, Liu P, et al. Influence of mortality on estimating the risk of kidney failure in people with stage 4 CKD. Journal of the American Society of Nephrology 2019;30:2219-27.
7. Sumida K, Nadkarni GN, Grams ME, et al. Conversion of Urine Protein-Creatinine Ratio or Urine Dipstick Protein to Urine Albumin-Creatinine Ratio for Use in Chronic Kidney Disease Screening and Prognosis : An Individual Participant-Based Meta-analysis. Ann Intern Med 2020;173:426-35.
8. Liu P, Quinn RR, Lam NN, et al. Accounting for Age in the Definition of Chronic Kidney Disease. JAMA Intern Med 2021;181:1359-66.
9. Vestergaard SV, Christiansen CF, Thomsen RW, Birn H, Heide-Jorgensen U. Identification of Patients with CKD in Medical Databases: A Comparison of Different Algorithms. Clin J Am Soc Nephrol 2021;16:543-51.
10. Inker LA, Eneanya ND, Coresh J, et al. New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. N Engl J Med 2021;385:1737-49.
11. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. Annals of internal medicine 2009;150:604-12.
12. Tonelli M, Vachharajani TJ, Wiebe N, et al. Methods for identifying 30 chronic conditions: application to administrative data. BMC Med Inform Decis Mak 2015;15:31.
13. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. Stat Med 2014;33:3191-203.
14. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999;18:2529-45.
15. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6.