Predicting the risks of kidney failure and death in adults with moderate to severe chronic kidney disease: multinational, longitudinal, population based, cohort study (KDpredict)

Ping Liu,[1] Simon Sawhney,[2] Uffe Heide-Jørgensen,[3] Robert Ross Quinn,[1] Simon Kok Jensen,[3] Andrew Mclean,[2] Christian Fynbo Christiansen,[3] Thomas Alexander Gerds,[4] Pietro Ravani[1]

1. Departments of Medicine and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
2. Aberdeen Centre for Health Data Science, University of Aberdeen, Scotland
3. Department of Clinical Epidemiology, Department of Clinical Medicine, Aarhus University and Aarhus University Hospital, Denmark
4. Department of Public Health, University of Copenhagen, Denmark

Supplementary Appendix 1

Contents

Detailed Methods 4
Ethics 4
Data sources 4
Key study dates 4
Sustained eGFR methodology for cohort formation and outcome definition 5
Predictors 5
Data differences across three cohorts 6
Sample size considerations 6
Missing data 6
Statistical analysis 6
Supplementary Tables 11
Table S1: Analysis plan 11
Table S2: Sustained eGFR methods for cohort formation and kidney failure outcome ascertainment 12
Table S3: Codes for identifying kidney transplantation using administrative data (Alberta) 13
Table S4: Codes for identifying comorbidities using administrative data 14
Table S5: Base learner design 15
Table S6: Cohort formation 16
Table S7: Summary of follow-up data in three cohorts 17
Table S8: Hazard ratios and 95% confidence intervals for kidney failure and death 18
Supplementary Figures 19
Figure S1: Distribution of study participants and estimated actual 5-year risks of kidney failure and death by CKD stage, albuminuria (Alberta cohorts) 19
Figure S2: Distribution of study participants and predicted 5-year risks of kidney failure and death by CKD stage, albuminuria (external cohorts) 20
Figure S3: Distribution of study participants and estimated actual 5-year risks of kidney failure and death by CKD stage, albuminuria (external cohorts) 21
Figure S4: Scatter plots of predicted risks of kidney failure and death at 2 and 5 years (Alberta cohorts) 22
Figure S5: Distribution of predicted mortality by predicted risk of kidney failure 23
Figure S6: Proportion of people above proposed mortality thresholds according to the risk of kidney failure 24
Figure S7: Agreement between individual 2- and 5-year risk predictions of kidney failure (4- vs 6-variable super-learner) 25
Figure S8: Calibration of 4- and 6-variable super-learner for 2- and 5-year prediction of kidney failure 26
Figure S9: Calibration of 4- and 6-variable super-learner for 1-, 3- and 4-year prediction of kidney failure 28
Figure S10: Calibration of 4- and 6-variable super-learner for 1- and 2-year prediction of kidney failure in the full G3bG4-CKD cohort 30
Figure S11: Agreement between individual 2- and 5-year risk predictions of kidney failure (original KFRE vs 4-variable super-learner) 32
Figure S12: Original and recalibrated KFRE vs 4-variable super-learner for 2- and 5-year prediction of kidney failure in people with G3bG4-CKD 33
Figure S13: Agreement between individual 2- and 5-year risk predictions of death (4- vs 6-variable super-learner) 35
Figure S14: Calibration of 4- vs 6-variable super-learner for 2- and 5-year prediction of death 36
Figure S15: Calibration of 4- and 6-variable super-learner for 1-, 3- and 4-year prediction of death 38

Figure S16: Temporal testing (kidney failure) 40
Figure S17: Temporal testing (mortality) 42
Figure S18: Agreement between individual 2- and 5-year risk predictions of kidney failure (eGFR calculated with EPI 2009 vs EPI 2021 formula) 44
Figure S19: Agreement between individual 2- and 5-year risk predictions of death (eGFR calculated with EPI 2009 vs EPI 2021 formula) 46
Figure S20: Decision curve analysis of mortality 48
Supplementary References 49

Detailed Methods

Ethics

The institutional review boards at the Universities of Alberta (Pro00053469) and Calgary (REB16-1575) approved this study with a waiver of participant consent. In Denmark, the study was reported for institutional registration, but ethical approval is not needed for registry-based research. Use of Grampian unconsented, pseudonymized, routinely collected health data were provided by Northwest Research Ethics Committee (19/NW/0552), Grampian Caldicott guardian, and the National Health Service (NHS) Research and Development.

Data sources

We used a comprehensive set of registry data to identify the population at risk and to define predictors and outcomes. All sites used person-level linked, population-based administrative health data that include demographics, vital statistics, laboratory data, ambulatory and inpatient datasets, physician claims.[1-4] The province of Alberta in Canada and Scotland have renal program repositories with dialysis or kidney transplant related records. Alberta, Scotland and Denmark offer unique opportunities for population-based studies, enabling: (1) whole population coverage; (2) longitudinal data over a long term; (3) rich lab data; (4) governance to enable secure federated analysis.[1-4]

Key study dates

Date range for cohort entry:
Alberta full cohort: April 1, 2008, to March 31, 2019 (**Table S1**)
Denmark cohort: January 1, 2007, to December 31, 2017, for the Central Denmark Region and January 1, 2012, to December 31, 2020, for the North Denmark Region.
Scotland cohort: January 1, 2011, to December 31, 2019.

Study end date:
Alberta cohorts: March 31, 2020.
Denmark cohort: December 31, 2018 for the Central Denmark Region and December 31, 2021 for the North Denmark Region.
Scotland cohort: December 31, 2020.

Look-back time window for eGFR screening (to exclude prevalent CKD patients):
Alberta cohorts: use prior eGFR data from May 1, 2002.
Denmark cohort: use prior available eGFR data from January 1, 1990.
Scotland cohort: use prior eGFR data from July 1, 2009.

Look-back time window for previous chronic dialysis or kidney transplant:
Alberta cohorts: use administrative data from April 1, 1994, and provincial registry data from January 1, 2001.
Denmark cohort: use surgery codes (kidney transplant) from January 1996, procedure codes (dialysis) from January 1999, and diagnosis codes indicating either of these, which date back to 1994 (when ICD-10 was introduced in Denmark).
Scotland cohort: all kidney replacement therapy recorded in renal information management system episodes captured back to January 1, 1972.

Look-back time window to define comorbidities:
Alberta cohorts: use administrative data from April 1, 1994.

Denmark cohort: use 5 years look-back for diagnosis codes and 1 year for prescriptions from cohort entry of the individual patient.
Scotland cohort: use Scottish Morbidity Record 01, back to January 1, 2004.

Sustained eGFR methodology for cohort formation and outcome definition

General approach. We screened all eGFR data in the laboratory repository of each jurisdiction. Criteria for cohort entry (incident G3bG4-CKD) and kidney failure (kidney outcome), were eGFR reduction below the thresholds of 45 and 10 mL/min/1.73m^2, respectively, sustained for >90 days (recommended chronicity criterion).[5-7] We calculated eGFR using the CKD-EPI-2009 formula[8] to identify participants and capture outcomes, excluding the race coefficient, with serum creatinine values standardized to isotope dilution mass spectrometry-traceable methods. We used 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 used the mean value of eGFR when there were multiple measurements on the same day. We also calculated the baseline eGFR from the index serum creatinine using the CKD-EPI-2021 race-free formula to compare the performance of models using the 2009 and 2021 formulas.[9]

Cohort formation. To identify people with sustained reduction of eGFR meeting criteria for cohort entry (G3bG4-CKD), we screened each individual's series of ≥2 consecutive eGFR tests where the first and last eGFR were separated by >90 days, the first and all possible intervening measurements within 90 days were <45 mL/min/1.73m^2 and the last eGFR was 15-44 mL/min/1.73m^2. We selected the earliest series (qualifying period) where all eGFR measurements met the eGFR requirement for cohort entry (**Table S2**). The date of the last eGFR in the qualifying period defined the index date (i.e., cohort entry or prediction time origin). We excluded people who had received chronic dialysis or kidney transplant or had had sustained eGFR <15 mL/min/1.73 m^2 for more than 90 days (stage 5 CKD) on or before cohort entry, because these people meet the criteria for kidney failure according to guidelines[10] and thus they are not at risk of kidney failure.

We included people who had records of urine albumin-to-creatinine ratio (ACR), urine protein-to-creatinine ratio (PCR), or urine dipstick in the 3 years preceding cohort entry, because albuminuria is a required input in the current kidney failure risk prediction tool.[11 12] We excluded individuals who had no proteinuria measurements within 3 years before cohort entry as the probability of a missing value for this variable is likely to be related to other predictors or unobserved variables,[13] and thus cannot be considered missing completely at random. A complete case analysis can be unbiased under missing at random or missing not at random.[14]

Kidney failure outcome. To capture kidney failure outcome defined by changes in eGFR, we used the same 'sustained' methodology (**Table S2**). Kidney failure was defined by the earliest of initiation of kidney transplant, chronic dialysis, or sustained eGFR <10 mL/min/1.73 m^2 for >90 days.

Predictors

At cohort entry, we considered age, sex, eGFR, ACR, history of diabetes, and cardiovascular disease (presence of congestive heart failure, myocardial infarction, peripheral vascular disease, or stroke or transient ischemic attack) for main analyses, and chronic pulmonary disease, and any cancer except malignant neoplasms of skin (**Table S3**), for descriptive purposes.[15]

When there were multiple same-type proteinuria measurements on the same day, we used the median for ACR or PCR and applied the floor function of median category for urine dipstick. We used the most recent outpatient proteinuria values within 3 years before study entry, with the following types of

measurement in descending order of preference: ACR, PCR, and urine dipstick. In our previous work we found that in Alberta 10-15% of people with newly documented CKD had no information on albuminuria or proteinuria in the 3 years before CKD diagnosis, and most people who received testing had a urine dipstick only.[16] In the selected external sites for model testing, urine dipstick results were not available. ACR can be calculated from urine dipstick although ACR calculated from dipstick or PCR vs. measured ACR may reduce model performance.[16] Nevertheless, we included people who had urine dipstick measures of proteinuria only (i.e., did not have PCR or ACR) in the Alberta cohort, because in Alberta, unlike Denmark and Scotland, dipsticks are part of usual care and workflow for the information management system. This minimizes the number of people excluded from the study, allowing the prediction tool to learn from more diverse, population-based data and enhances its usability in settings where urine dipstick tests are available. We used the crude formula proposed by Sumida et al to obtain ACR from PCR or dipstick.[17]

Data differences across three cohorts

For Alberta, Denmark, and Scotland cohorts, there was no differences in eligibility criteria for study population, outcome definition, and definitions for predictors. However, we allowed for regional variations in clinical practice and data availability or structure to identify comorbidities using administrative data. Specifically, Denmark and Scotland used additional information on medication to identify a history of diabetes or asthma (**Table S4).**

Sample size considerations

An appropriate setup for a sample size calculation requires information on data quality, predictors, and the statistical modeling approach. Existing research on this subject is typically focused on a fully specified regression model.[18] However, rules of thumb (e.g., 10 events per model degree of freedom) and other ad hoc formulas are too simple, because they do not address the process of finding the best prediction model, using, for example, meta-algorithms designed to learn which predictor variables should be included and how (super-learner). Using the approach recommended by Riley et al.[18] (package `pmsampsize` in R) and inputs from our previous studies (kidney failure rate of 1 per 100 patient-years, max parameters N = 20 for the 6-variable model, mean follow-up of 5 years)[5 6] and an R^2 of 0.05, the minimum sample size required for new model development is 3500 (this is about the size of the original KFRE study[11]). However, we acknowledge that the best approach to power and sample size calculation in this area is to simulate data with the computer using different alternative sample sizes and some criteria for the desired population average prediction performance. Of note, the population-based learning cohort was about 15 times larger than the cohort used to develop the KFRE.[11]

Missing data

The Alberta cohort excluded 12% of individuals who met eGFR criteria for entry but had no records of outpatient albuminuria in the preceding 3 years (albuminuria is a well-known predictor of adverse outcomes in people with CKD).[19] We performed a complete-case analysis, because albuminuria is a required input in the current benchmark prediction tool, KFRE,[11 12] and people without information on albuminuria are very different from those who receive an albuminuria testing. They are older, more likely to have cancer and other comorbidities, and less likely to be referred to a nephrologist.

Statistical analysis

We used standard methods for descriptive statistics. We used the reverse Kaplan-Meier estimator for the censoring distribution to summarize the follow-up time. We summarized the cumulative incidence of kidney failure using the Aalen-Johansen estimator and all-cause mortality risk using the Kaplan-

Meier estimator. Since these two estimators assume that the censoring distribution is independent of the distribution of the predictors, we presented stratified predicted absolute risks in main analyses and estimated actual risks in secondary analyses.

Super-learner strategy. There are different strategies for learning medical risk prediction models from data ('learners'), and it is impossible to anticipate which of them is the most suitable for a given prediction task. In fact, there are many regression models, and each can be specified in different ways to handle interactions or non-linear effects. Also, many machine learning algorithms for risk prediction exist, and can be tuned in different ways to configure the learning process. The super-learner is a meta-algorithm that alleviates these model selection concerns by providing the freedom to consider many alternative prediction model algorithms and either combine them in an ensemble (ensemble super-learner) or select the best performing one among them (discrete super-learner). The pre-specified set of prediction model algorithms (e.g., recommended by collaborators or subject-matter experts) is called 'library'.[20]

To create the library of prediction model algorithms (base learners), one author (PL) generated synthetic data from older Alberta data (cohort entry between April 1, 20081 and March 31, 2011), which were used by two other authors for blind algorithm design (TAG, PR). The synthetic data were used to design and test the regression models and the random forest algorithms with different configurations (different values for the tuning hyperparameters) to ensure that they would run without error in supervised learning (using the original individual-level data). We had no access to any original individual-level data during the design phase of the study and had to write all codes in advance. The synthetic sample had the same probability distribution of the combinations of the predictor variables from older Alberta data (cohort entry from April 1, 2008, to March 31, 2011) and time to event altered with random numbers. We designed different regression models, considering clinical judgement,[10] and machine learning algorithms with different configurations (base learners).[21] There were two sets of libraries of base learners: one for kidney failure that accounted for the competing risk of death (cause-specific Cox models[22] and random forest for competing risks)[23] and one for time to death analysis (standard Cox models and random survival forest algorithms).[24] These libraries of base learners (**Table S5**) included regression models with different settings (coefficients or internal parameters that a model learns from the data), and random forest algorithms with a range of hyper-parameters (external parameters that control the learning process).

For regression models we considered different transformations including restricted cubic splines of 0-3 continuous variables (age, eGFR, log-ACR), and first order interactions between predictors based on clinical judgement and existing studies[10] (expert-derived or clinical modeling culture).[21] Spline knots were either estimated based on variable distribution or prespecified (age: 65, 75, and 85 years; eGFR: 25, 35, and 40 mL/min/1.73 m^2; log-ACR: 3.4, 5.7, 6.3, corresponding to 30, 300 and 600 mg/g). For random forest tuning, we considered a range of values for the hyperparameters. Hyperparameter optimization of a random forest (tuning) consists of finding values for number of trees, random split points for continuous variables, minimal terminal node size, and number of candidate features for random bootstrap sampling tried at each node split that minimises out-of-bag error. Random forests are free from assumptions of hazard proportionality, linearity of effects, and absence of interactions. Inferior performance of a random survival forest in head-to-head comparison with a rival semi-parametric Cox model provides indirect evidence of the goodness of fit of the semi-parametric model.[25]

We constructed base learners for two sets of predictor variables: (1) age, sex, eGFR, ACR and (2) with the same four variables plus diabetes and cardiovascular disease summarized as a binary variable. While additional variables may increase prediction accuracy, comorbidity variables may also be

recorded in different ways across jurisdictions. Use of additional variables may alter model performance for reasons unrelated to the model and may reduce its usability (see Protocol). For each set we prepared a version including eGFR calculated using the 2021 formula and one with the 2009 formula.[9 26]

In supervised learning (outcome analysis), we used contemporary Alberta data (cohort entry between April 1, 2011 and March 31, 2019) to identify the strongest learner by fitting a discrete super-learner ('one winner takes all strategy') with each library of base learners that was specified in unsupervised learning using synthetic data (**Table S5**).[20] A super-learner is a machine learning algorithm that uses cross-validation to train a series of base learners and evaluate their prediction performance. The super-learner used internal cross-validation based on 500 bootstrap sets each obtained by random subsampling 63.2% of the training cohort for learning (in-bag) and 36.8% to calculate the prediction performance (out-of-bag). We used the leave-one-out bootstrap for averaging the performance results across multiple splits.[27] The super-learner could include a variable number of winners per each set of predictors (4- and 6-variable sets), from one winner per outcome (kidney failure for the competing risk library and death for the survival library) in case the same model outperforms all the other rival models for that outcome at all time horizons (years 1-5) to five winners (one per each prediction time horizon) per each outcome. The discrete super-learner selected the outcome-specific learner with lowest mean of the five Brier scores. From each version of the super-learner (4-, 6- variable and with eGFR calculated using the CKD-EPI-2009 and CKD-EPI-2021 formulas), we obtained the two outcome-specific winners that had the lowest mean Brier score over all time horizons (years 1-5), provided that the winners were all cause-specific (for kidney failure) or standard Cox models (for death). If the super-learner included a random forest for any time horizon, we planned to use that random forest for predictions at that time horizon and the Cox model with the lowest mean Brier score for the other time horizons.

Transportability (geographical testing). To investigate to what extent the super-learner trained in Alberta could be exported to other jurisdictions, the external testing teams (in Denmark and Scotland) compared the 4-variable super-learner to the current benchmark model (4-variable kidney failure risk equation, KFRE) for 2- and 5-year kidney failure risk predictions (the only time horizons the KFRE considers).[12] Since prediction time horizons of interest depend on disease severity (i.e., 2-year predictions for specialist or advanced care planning for kidney failure are only of interest to people with more severe CKD), 1-2-year kidney failure risk predictions were evaluated only in people with G4-CKD in main analyses. In secondary analyses we included also people with G3b in short-term prediction of kidney failure. Mortality risk predictions were assessed at years 1-5 for all participants. Different formulations of the super-learner (4- and 6-variable) were also evaluated for 1-5-year predictions of both risks. These analyses used the full set of each external cohort without cross-validation.

Model performance measures. We used predicted risk scatterplots to visualize potential differences (disagreement) between individualized risk predictions from rival models. We assessed calibration in the small using histogram-type plots, time-dependent area under the receiver operating characteristic curve (AUC, a measure of discrimination, a ranking statistic; the higher the better),[28] Brier score (prediction error, a measure of both calibration and discrimination; the lower the better),[29] and index of prediction accuracy (IPA, a measure of average prediction performance; the higher the better).[30] Note that among these performance measures, the only strictly proper scoring rule is the Brier score (from which the IPA is derived, IPA = 1 – Brier[model]/Brier[null]).

For calibration, we used histogram-type calibration plots to purposefully overrepresent low-risk categories, based on existing studies showing that most people with CKD have a predicted 5-year risk of kidney failure <10%.[5] We categorized the predicted risks from each model for all individuals in each testing dataset into a pre-specified number of equally large groups (tenths of predicted risk, i.e., 10 groups defined by cut-off points corresponding to the deciles of the predicted risk distribution of the test dataset) and compared in each group the mean predicted risk to the Aalen-Johansen estimated actual risk. The number of bins was pre-specified to prevent analyst manipulation. A model is well calibrated if the estimated and the actual risks are similar in all groups. However, calibration plots are not always easy to read and do not provide a simple answer as to which of two rival models is more accurate. In fact, the predicted risk quantiles from which the bars of the calibration plots are generated differ by model. The numeric Brier score is a preferable performance measure for model ranking.

We used the inverse probability of censoring weighted estimates of the AUC to estimate model discrimination (values of 50% or lower indicate the model is useless or harmful).[28] Being a ranking statistic, AUC cannot stand alone to indicate the model has value (a well-discriminating model may have poor calibration). We did not consider specific values of AUC, but estimated AUC to ensure values were >50%.

The Brier score, which depends on both discrimination and calibration,[31] is the average of the squared differences between the predicted risks and estimated actual outcomes (i.e., the mean squared error of prediction), ranging from 0 (perfect model) to 25% (non-informative model). We did not test for differences in Brier score; instead, we used the Brier score as a measure of loss function for model ranking. Because the values of the Brier score are on a scale that does not allow unconditional interpretation (i.e., requires the knowledge of the Brier score of the benchmark null model), we used the IPA to quantify the overall accuracy of each model (the higher IPA the better). The IPA, which is one minus the ratio of the model Brier score to the null model Brier score and thus depends on both discrimination and calibration,[32] ranges from <0 (harmful model), to 0 (useless model) to >0 (useful model).

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 considered clinically relevant changes (>10% difference) in individualized predicted risks, calibration and absolute values of the Brier score in model evaluation. We also presented 95% confidence intervals of all scores, but avoided formal comparisons.

In main analyses, the rival model comparisons for geographical testing included:

1) 4-variable vs 6-variable super-learner, for kidney failure and all-cause death, at 1-5 years (G4 only for kidney failure at 1-2 years);
2) Super-learner (4-variable) vs KFRE for kidney failure, at 2 (G4 only) and 5 years (full cohort)

No existing models were found to compare the performance of the super-learner vs a benchmark for the outcome of death (the Grams study[33] uses race and blood pressure values among predictor variables and was developed in G4-CKD and G5-CKD (Guidelines define G5-CKD as kidney failure[10]).

Temporal testing of retrained models. Ideally, retrained models (or retrained super-learner, potentially including new models or algorithms, or new predictors) should be tested on unseen, future data in each external site as new data become available (temporal testing). Retraining may be required over time to account for temporal changes in clinical practice, health policy or population characteristics within a jurisdiction. To evaluate if the super-learner models will perform well on future patients, we retrained

them using older Alberta data (index date between April 1, 2008 and December 31, 2014) and presented their cross-validated performance (their performance on the training set using cross-validation) along with their performance on the full set of temporally distinct, more recent data without cross-validation (index date between January 1, 2015 and March 31, 2019). By splitting the data into independent training and testing sets, cross-validation tests an average model and simulates how well the model will perform when challenged with unseen, future data. Given the relatively small size of the Denmark and Scotland cohorts, this was possible only in the Alberta cohort.

Clinical usefulness. To illustrate the clinical value of the super-learner, we plotted the predicted risks of kidney failure and all-cause death for hypothetical individuals with characteristics associated with combinations of high/low risk for kidney failure and high/low risk for death. We obtained 95% confidence intervals from bootstrap distributions (1000 replicates). We translated the super-learner into an online calculator to enhance clinical usability.

Other analyses. We used scatterplots to visualize potential differences between individual risk predictions from the super-learner using the 2021 formula instead of the 2009 formula to calculated eGFR. Decision curve analysis (DCA) was used to assess the clinical utility of different models, by plotting the "net benefit" against pre-specified "threshold probabilities".[34] Net benefit measures the trade-off between true positives and false positives in a prediction model at different threshold probabilities. It is a sum of true-positive minus false-positive predictions weighted by the threshold probability. The threshold is a clinically derived value that varies depending on how risk averse the decision-maker is. It is a value where the end-user would be satisfied with the trade-off between the harm of delaying an intervention (e.g., unplanned treatment for kidney failure) and unnecessary intervention (unnecessary planning for kidney failure treatment). For example, a 2-year risk threshold of 10% implies that for every 1 true-positive case identified by the decision strategy, fewer than 9 false positives would be an acceptable trade-off. This means weighting the finding a high-risk patient as 9 times more important than avoiding unnecessary referral for enhanced care. Although the absolute value of net benefit is an abstract concept, the higher the positive value of net benefit, the greater the clinical value of a prediction tool, allowing a comparison of different strategies for clinical decision making.

We used the packages `riskRegression` and `randomForestSRC` in R.

Supplementary Tables

Table S1: Analysis plan

Step	Data used	Objective and methods
Data preparation	Routinely recorded health data (Alberta); cohort creation by a team with access to individual-level-data	Creation of two cohorts in Alberta, Canada: N.1 full cohort (cohort entry 2008-04-01 to 2019-03-31, follow-up to 2020-03-31) N.2 older cohort (cohort entry 2008-04-01 to 2011-03-31) for synthetic data creation N.3 synthetic data (obtained from cohort N.2) with the same probability distribution of the combinations of the predictors as the cohort N.2 but outcome altered with random numbers N.4 contemporary cohort for supervised learning (cohort entry after 2011-03-31)
Library of learners	Synthetic data, N.3. Analyses conducted by researchers blinded to the individual-level-data	Creation of 4 libraries of 'base learners' per outcome (one with 4 and one with 6 predictors for each eGFR calculation method). Learners were different models (Cox models with different settings, i.e., internal parameters that the models 'learn' from the data) and machine learning algorithms (random forests with different hyper-parameters, i.e., external parameters that the modeller tunes to control the learning process)
Supervised learning	Original, contemporary Alberta cohort N.4 (cohort entry after 2011-03-31). Conducted by researchers who had access to individual-level-data	Identification of the strongest learners by fitting a discrete super-learner with each library of learners. A super-learner is a machine learning algorithm that uses cross-validation for training and testing each learner performance. For each library, the super-learner identified the outcome-specific learner with the lowest cross-validated time-dependent Brier score at 1-, 2-, 3-, 4-, and 5 years, and then selected the outcome-specific learner with lowest mean of the 5 scores (winner). The super-learner used cross-validation based on 500 bootstrap sets each obtained by random subsampling 63.2% of the cohort for learning (in-bag) and 36.8% for testing (out-of-bag). The leave-one-out bootstrap was used for averaging the performance results across multiple splits.
Transportability assessment	Full set of the Danish and Scottish data, without cross-validation	Evaluation of the performance of the super-learner trained in Alberta (cohort N.4) when used in different settings (previously unseen data). Comparison with the existing benchmark for kidney failure risk prediction
Temporal testing	Routinely recorded health data (Alberta)	N.5 older data for model retraining (cohort entry until 2014-12-31) N.6 recent data for temporal testing of retrained models (cohort entry after 2014-12-31)

Note: Researchers who had access to individual-level data created all cohorts for analyses (data preparation) and completed supervised learning. Researchers who were blinded to individual-level data conducted unsupervised learning and super-learner library design.

Table S2: Sustained eGFR methods for cohort formation and kidney failure outcome ascertainment

	Qualifying period (>90 days)		
Cohort	*First eGFR*	*Intervening eGFR*	*Last eGFR (index date)*
G3bG4-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.[26]

- G3bG4-CKD refers to GFR categories of CKD. G3b-CKD: CKD with moderately to severely decreased GFR (eGFR 30-44 mL/min/1.73 m^2); G4-CKD: CKD with severely decreased GFR (eGFR 15-29 mL/min/1.73 m^2).
- 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 (and the index eGFR ≥15 mL/min/1.73 m^2).
- 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.
- We excluded people who had received chronic dialysis or kidney transplant or had had sustained eGFR <15 mL/min/1.73 m^2 for more than 90 days (stage 5 CKD) on or before cohort entry, because these people meet the criteria for kidney failure according to guidelines[10] and thus they are not at risk of kidney failure.
- Main kidney failure (KF) outcome was defined as the earliest of initiation of kidney transplant, chronic dialysis or sustained eGFR <10 mL/min/1.73 m^2 for >90 days. The outcome date of kidney failure according to eGFR criteria was the date of last eGFR during the qualifying period for kidney failure defined based on eGFR.

Table S3: Codes for identifying kidney transplantation using administrative data (Alberta)

1) Physician claims: Alberta Health Care Insurance Plan, Medical Procedure codes

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

2) Hospitalizations: Canadian Classification of Diagnostic, Therapeutic, and Surgical Procedures codes

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 S4: 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 cohort entry for diagnoses	140-209, except 173	C00-C96, except C44	-

Legend: We used a broader definition for cancer (presence of any cancer except malignant neoplasms of skin). We identified all other comorbidities using validated algorithms.[35] Because these comorbidities are considered as permanent conditions, Alberta and Scotland cohorts used all available prior data from each jurisdiction to identify baseline comorbidities, except cancer requiring 5-year look-back from cohort entry for diagnoses. The Denmark cohort used a 5-year look-back for all diagnosis codes (not just "most responsible") and 1-year look-back for prescriptions.

Look-back window:

Alberta cohort: using administrative data from April 1, 1994 to cohort entry, which was between April 1, 2008 and March 31, 2019.

Denmark cohort: use 5 years look-back for diagnosis codes, which was back to 1994.

Scotland cohort: use Scottish Morbidity Record 01, back to January 1, 2004.

Table S5: Base learner design

*Base learners	#N of variables	Strata (stage)	^Splines	Interactions	N trees	Features tried at each node	Split N	Node size
Cox*	4	no	From 0 to 3 (age, eGFR, log-ACR)	From 0 to 2 (sex*age, stage*log-ACR)	-	-	-	-
		yes	From 0 to 2 (age, log-ACR)	From 0 to 1 (sex*age); with/without stratum per term interactions	-	-	-	-
	6	no	From 0 to 3 (age, eGFR, log-ACR)	From 0 to 8 (sex*age, DM*age, CV*age, CV*DM, stage*log-ACR, DM*log-ACR)	-	-	-	-
		yes	From 0 to 2 (age, log-ACR)	From 0 to 7 (sex*age, DM*age, CV*age, CV*DM, DM*log-ACR); with/without stratum per term interactions	-	-	-	-
RSF	4	-	-	-	100-200	2-3	10	10
	6	-	-	-	100-200	3-4	10	10

Legend:
*Base learners: CSC: cause-specific Cox for kidney failure; standard Cox models for death; RSF: random survival forest for competing risks or for death outcome
#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 (DM), any cardiovascular disease (CV)
Each base-learner library included many base learners characterized by different settings for the regression models (different combinations of stratification, spline number, and interactions) and different hyperparameters for the random survival forest algorithms. There were 4 libraries for each outcome, with 4 or 6 predictors, and with eGFR calculated with the 2009 EPI formula or the 2021 EPI formula.
^Splines: restricted cubic splines, with estimated or pre-specified knots (see supplemental methods): 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 S6: Cohort formation

	Alberta	Denmark	Scotland
Number of residents registered or with at least one health episode record, from 1 Apr 1994 to 31 Mar 2019 (Alberta), from 1 Apr 1994 to 31 Dec 2021 (Denmark), and from 1 July 2009 to 8 Mar 2020 (Scotland)	5,217,391	2,981,482	498,490
Exclusion criteria (N excluded)	5,149,449	2,963,954	490,750
No serum creatinine measurements, from 1 May 2002 to 31 Mar 2019 (Alberta), from 1 Jan 1990 to 31 Dec 2021 (Denmark), and from 1 Jul 2009 to 8 Mar 2020 (Scotland)	1,553,261	793,795	32,859
No outpatient serum creatinine measurements	205,000	149,420	10,342
All outpatient serum creatinine were tested under 18 years old	27,133	149,770	0[b]
All outpatient serum creatinine measurements were tested after the earliest of out-migration, registration end or accrual end (Alberta: 31 Mar 2019, Central Denmark Region: 31 Dec 2017, North Denmark Region: 31 Dec 2020, and Scotland: 31 Dec 2019)	6,268	261,440	518
Never had an outpatient eGFR <45 mL/min/1.73 m^2	3,180,035	1,433,910	417,446
Only 1 outpatient eGFR<45 mL/min/1.73 m^2	66,975	47,775	10,688
Did not meet the eGFR criteria for sustained G3bG4-CKD (see Table S2)	53,967	40,195	9,450
Index date was not between 1 Apr 2008 and 31 Mar 2019 (Alberta), 1 Jan 2007 and 31 Dec 2017 (Central Denmark Region), 1 Jan 2012 and 31 Dec 2020 (North Denmark Region), and 1 Jan 2011 and 31 Dec 2019 (Scotland)	46,611	57,470	6,579
Index date was on the earliest date of death, out-migration, registration end or accrual end	5	5[a]	15
Initiated chronic dialysis or received kidney transplant on or prior to cohort entry (look-back window for Alberta cohorts: administrative data from April 1, 1994 and provincial registry data from January 1, 2001; Denmark cohort: administrative data from 1994; Scotland cohort: renal information management system data from 1972)	1,042	630	0[c]
Sustained eGFR<15 mL/min/1.73 m^2 for >90 days on or before cohort entry	104	40	5
No ACR, PCR or urine dipstick protein measurements in the preceding 3 years	9,048	29,510	2,848
Study cohort (N)	**67,942**	**17,528**	**7,740**

a. Numbers in each step of the exclusion sequence are rounded to nearest 5 to prevent reporting numbers less than 5.
b. Access to blood tests of people aged <18 years was not covered by the ethics permissions in Scotland.
c. Zero exclusion for this step in Scotland due to prior exclusion from source dataset of any serum creatinine data after the date of kidney replacement therapy.

Table S7: Summary of follow-up data in three cohorts

	Alberta	Denmark	Scotland
N	67,942	17,528	7,740
No. of participants who developed kidney failure by year 5	1961	514	193
No. of participants who initiated maintenance KRT by year 5	1823	441	172
No. of participants died by year 5	20,895	5,697	2,738
Censoring distribution, median (IQR) time in years[a]	5.80 (3.31-8.64)	5.59 (3.53-7.85)	4.97 (2.78-7.41)
Rate of kidney failure (per 100 person-year), 95% CI	1.11 (1.07-1.15)	0.95 (0.88-1.02)	0.82 (0.71-0.93)
Rate of death (per 100 person-year), 95% CI	9.52 (9.41-9.63)	9.77 (9.54-9.996)	11.82 (11.41-12.22)

Legend: IQR = inter-quartile range; CI = confidence interval. KRT = kidney replacement therapy. Kidney failure (main outcome) was defined as the earliest of initiation of kidney transplant, chronic dialysis or sustained eGFR <10 mL/min/1.73 m^2 for >90 days.

a. We used the reverse Kaplan-Meier estimator for the censoring distribution to summarize the follow-up time.

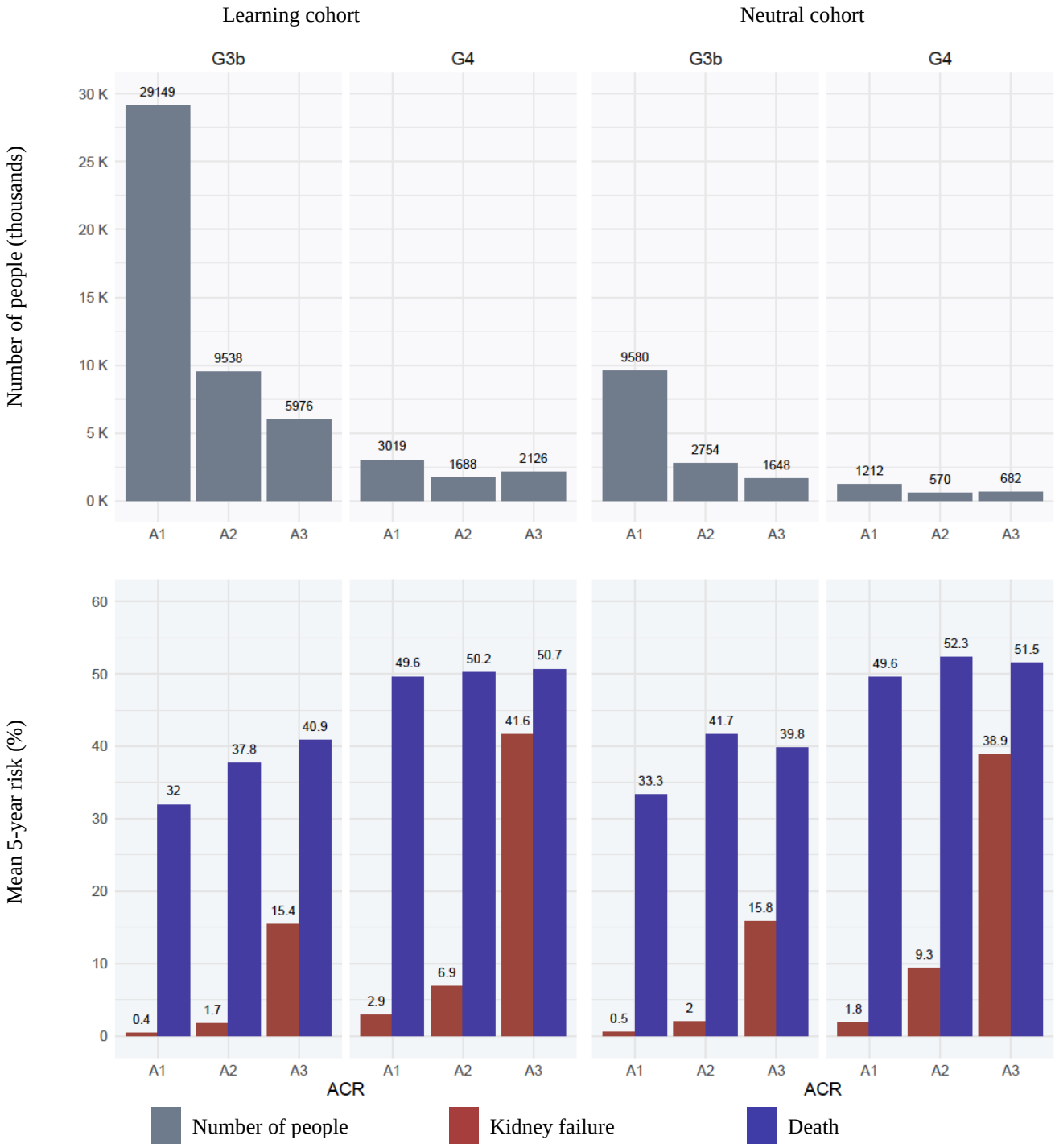
Table S8: Hazard ratios and 95% confidence intervals for kidney failure and death

	CKD EPI 2009		CKD EPI 2021	
	Kidney failure	**Death**	**Kidney failure**	**Death**
4-variable super-learner				
Male	1.31 (1.16 −1.49)	0.47 (0.27 −0.80)	1.39 (1.23 −1.57)	0.50 (0.30 −0.83)
Age (year)	0.97 (0.96 − 0.97)	1.03 (1.02 − 1.03)	0.97 (0.96 − 0.97)	1.03 (1.02 − 1.03)
Age (spline)	0.93 (0.91 − 0.96)	1.04 (1.04 − 1.05)	0.93 (0.91 − 0.95)	1.04 (1.03 − 1.05)
eGFR	0.92 (0.91 − 0.93)	0.96 (0.96 − 0.97)	0.92 (0.91 − 0.93)	0.96 (0.96 − 0.97)
log ACR (mg/g)	1.71 (1.57 − 1.87)	1.17 (1.16 − 1.18)	1.68 (1.54 − 1.82)	1.16 (1.15 − 1.18)
log ACR (spline)	1.06 (1.00 − 1.12)	-	1.07 (1.01 − 1.13)	-
G4*male	0.88 (0.73 − 1.05)	1.68 (0.65 − 4.35)	0.80 (0.67 − 0.97)	1.70 (0.62 − 4.67)
G4*age	1.00 (1.00 − 1.01)	1.00 (0.99 − 1.01)	1.00 (1.00 − 1.01)	1.00 (0.99 − 1.01)
G4*age spline	1.01 (0.98 − 1.05)	0.98 (0.97 − 0.99)	1.03 (0.99 − 1.06)	0.98 (0.97 − 0.99)
G4*eGFR	0.99 (0.97 − 1.01)	1.01 (1.00 − 1.02)	1.00 (0.98 − 1.02)	1.01 (1.00 − 1.02)
G4*log ACR	0.78 (0.68 − 0.90)	0.94 (0.92 − 0.96)	0.79 (0.69 − 0.92)	0.95 (0.92 − 0.97)
G4*log ACR spline	1.08 (0.99 − 1.18)	-	1.07 (0.98 − 1.17)	-
Male*age	-	1.01 (1.01 − 1.02)	-	1.01 (1.01 − 1.02)
Male*age spline	-	0.99 (0.98 – 1.00)	-	0.99 (0.98 – 1.00)
G4*male*age	-	0.99 (0.98 – 1.00)	-	0.99 (0.98 − 1.01)
G4*male*age spline	-	1.01 (1.00 − 1.03)	-	1.01 (0.99 − 1.03)
6-variable super-learner				
Male	1.28 (1.13 − 1.46)	1.23 (1.19 − 1.27)	1.35 (1.20 − 1.53)	1.25 (1.21 − 1.29)
Age (year)	0.96 (0.95 − 0.96)	1.03 (1.03 − 1.04)	0.96 (0.96 − 0.97)	1.03 (1.03 − 1.04)
Age (spline)	0.95 (0.93 − 0.97)	1.04 (1.04 − 1.05)	0.94 (0.92 − 0.96)	1.04 (1.04 − 1.05)
eGFR	0.92 (0.91 − 0.93)	0.96 (0.96 − 0.97)	0.92 (0.91 − 0.93)	0.97 (0.96 − 0.97)
Diabetes	2.42 (1.20 − 4.89)	1.09 (1.05 − 1.12)	2.60 (1.32 − 5.10)	1.08 (1.05 − 1.12)
log ACR (mg/g)	1.86 (1.62 − 2.13)	1.15 (1.14 − 1.16)	1.85 (1.62 − 2.12)	1.15 (1.13 − 1.16)
log ACR (spline)	0.94 (0.86 − 1.04)	-	0.94 (0.85 − 1.03)	-
CVD	1.09 (0.96 − 1.24)	4.54 (2.62 − 7.85)	1.09 (0.96 − 1.23)	4.52 (2.67 − 7.64)
Diabetes*log ACR	0.85 (0.71 − 1.01)	-	0.83 (0.70 − 0.98)	-
Diabetes*log ACR spline	1.18 (1.04 − 1.33)	-	1.20 (1.07 − 1.35)	-
G4*male	0.89 (0.74 − 1.07)	0.97 (0.89 − 1.05)	0.82 (0.69 − 0.99)	0.95 (0.86 − 1.04)
G4*age	1.01 (1.00 − 1.01)	0.99 (0.98 – 1.00)	1.01 (1.00 − 1.01)	0.99 (0.98 – 1.00)
G4*age spline	1.01 (0.98 − 1.04)	0.99 (0.97 – 1.00)	1.02 (0.99 − 1.06)	0.99 (0.98 − 1.01)
G4*eGFR	0.99 (0.97 − 1.01)	1.00 (0.99 − 1.01)	0.99 (0.97 − 1.01)	1.00 (0.99 − 1.02)
G4*diabetes	0.28 (0.09 − 0.89)	1.00 (0.92 − 1.08)	0.20 (0.06 − 0.65)	1.03 (0.93 − 1.13)
G4*log ACR	0.67 (0.55 − 0.83)	0.95 (0.93 − 0.98)	0.64 (0.52 − 0.80)	0.95 (0.93 − 0.98)
G4*log ACR spline	1.20 (1.04 − 1.39)	-	1.25 (1.08 − 1.44)	-
G4*CVD	0.88 (0.73 − 1.06)	0.32 (0.12 − 0.83)	0.86 (0.71 − 1.04)	0.24 (0.09 − 0.68)
G4*diabetes*log ACR	1.36 (1.03 − 1.80)	-	1.51 (1.13 − 2.01)	-
G4*diabetes*log ACR spline	0.82 (0.68 − 0.98)	-	0.76 (0.63 − 0.92)	-
G4*CVD*age	-	1.01 (1.00 − 1.03)		1.02 (1.00 − 1.03)
G4*CVD*age spline	-	0.99 (0.98 − 1.01)		0.99 (0.97 − 1.01)
CVD*age	-	0.99 (0.98 – 1.00)		0.99 (0.98 – 1.00)
CVD*age spline	-	0.99 (0.99 – 1.00)		1.00 (0.99 – 1.00)

Legend: One version of the super-learner included eGFR calculated with the 2009 EPI creatinine-based formula, and one included the 2021 EPI creatinine-based race-free formula. Hazard ratios (95% CIs) were estimated from cause-specific Cox model for kidney failure and standard Cox model for death. CKD-EPI = chronic kidney disease epidemiology collaboration; eGFR = estimated glomerular filtration rate, in mL/min/1.73 m^2; ACR: measured albumin-to-creatinine ratio (ACR) or ACR calculated from protein-to-creatinine ratio (PCR) or urine dipstick. Conversion factor for ACR: 1 mg/mmol = 0.113 mg/g. CVD = congestive heart failure, myocardial infarction, peripheral vascular disease, or stroke or transient ischemic attack; G4 = severe chronic kidney disease (index eGFR 15-29 mL/min/1.73 m^2).

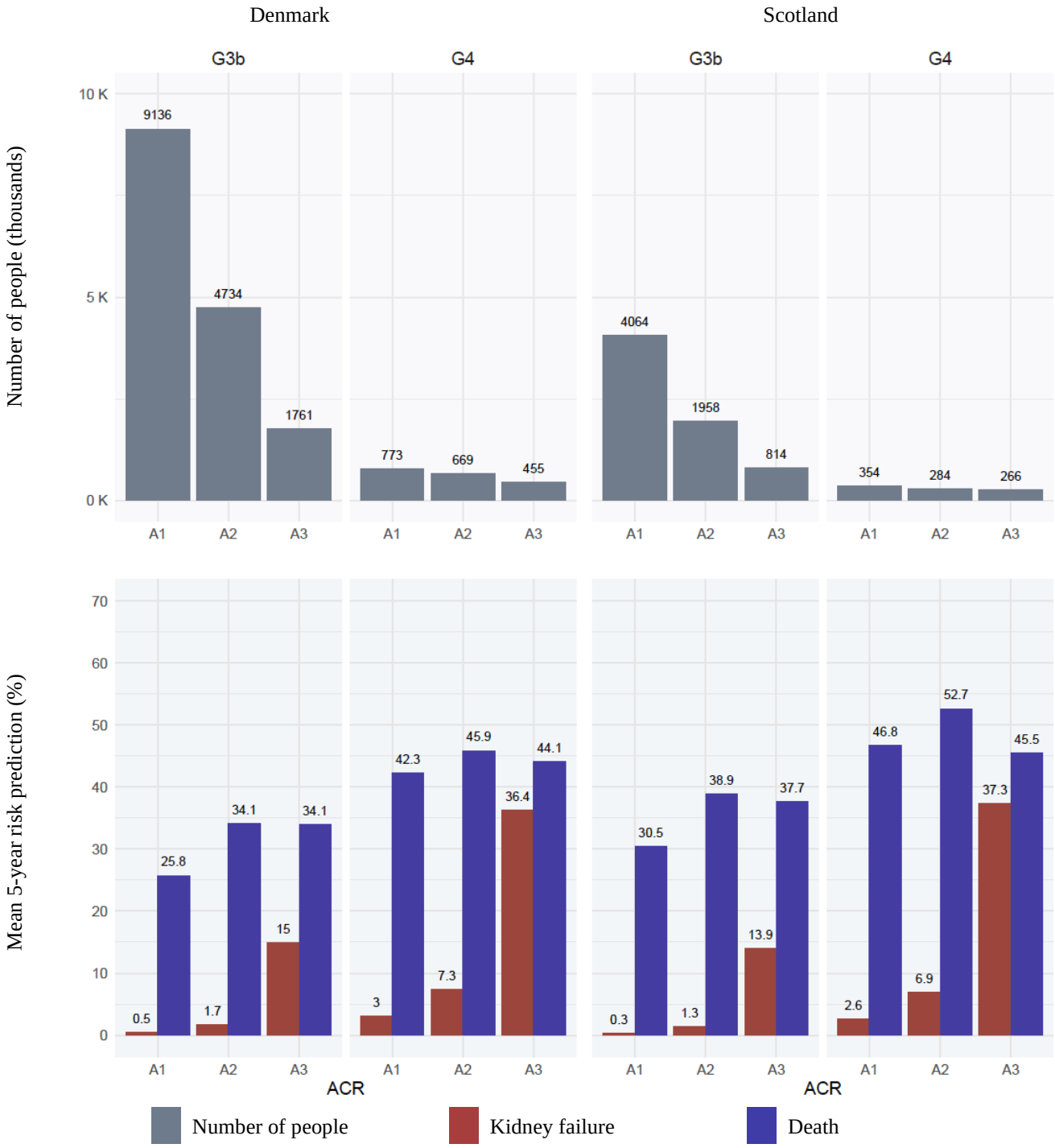
Supplementary Figures

Figure S1: Distribution of study participants and estimated actual 5-year risks of kidney failure and death by CKD stage, albuminuria (Alberta cohorts)



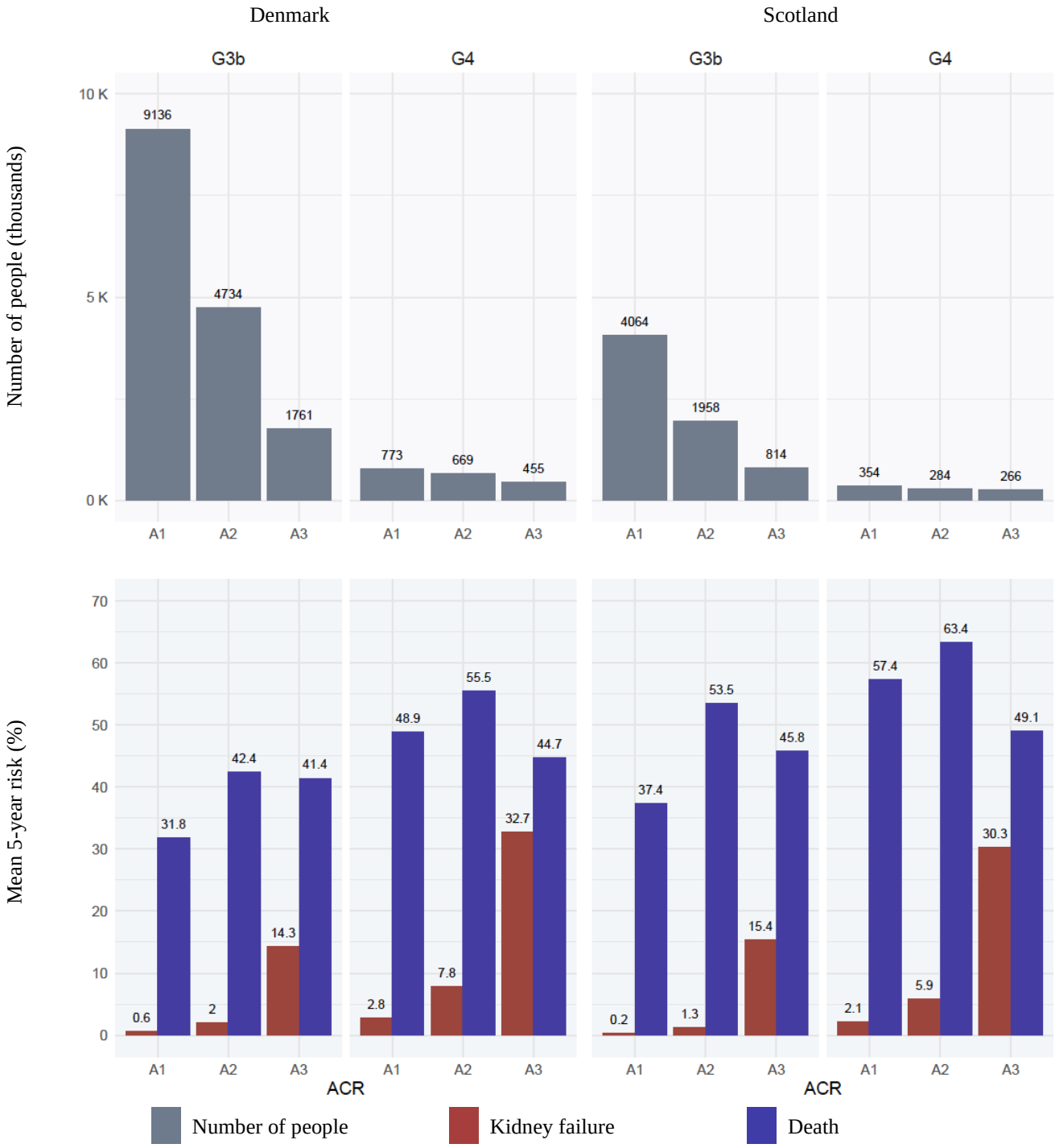
Legend: Data are from Alberta cohort data used for supervised learning (learning cohort, left panels, N=51,496) and to generate synthetic data for unsupervised learning (neutral cohort that the model has not seen, right panels, N=16,446). G3b indicates moderate chronic kidney disease (eGFR 30-44 mL/min/1.73 m^2); G4 indicates severe chronic kidney disease (eGFR 15-29 mL/min/1.73 m^2); ACR indicates albumin-to-creatinine ratio (A1 <30 mg/g, A2 30-300 mg/g, A3 >300 mg/g). Absolute frequencies (top panels) refer to number of people; percentages (bottom panels) refer to 5-year estimated actual risks of kidney failure and death.

Figure S2: Distribution of study participants and predicted 5-year risks of kidney failure and death by CKD stage, albuminuria (external cohorts)



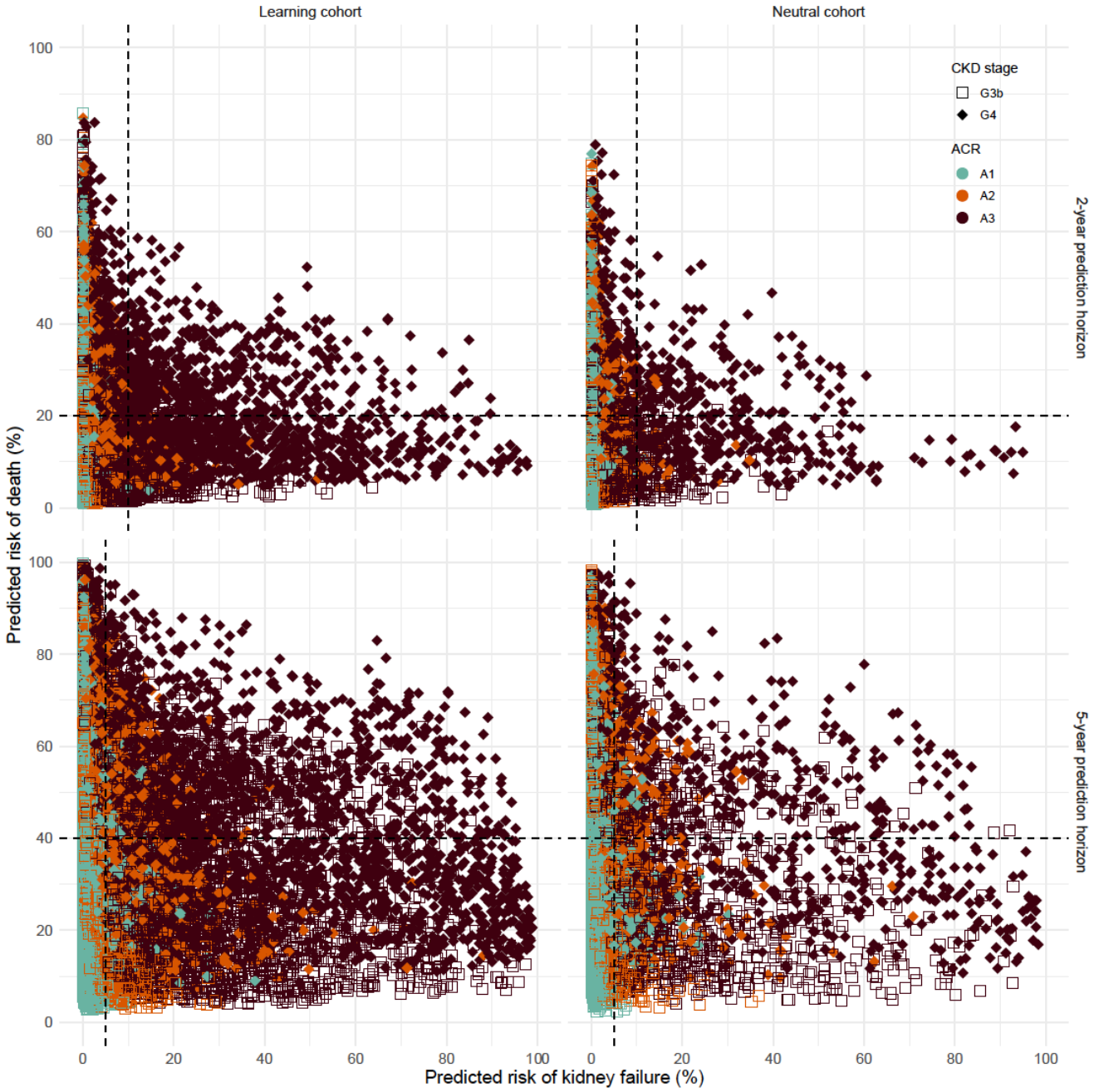
Legend: Data are from Denmark (left; N=17,528) and Scotland (right; N=7,740). G3b indicates moderate chronic kidney disease (eGFR 30-44 mL/min/1.73 m^2); G4 indicates severe chronic kidney disease (eGFR 15-29 mL/min/1.73 m^2); ACR indicates albumin-to-creatinine ratio (A1 <30 mg/g, A2 30-300 mg/g, A3 >300 mg/g). Absolute frequencies (top panels) refer to number of people; percentages (bottom panels) refer to 5-year predicted risks kidney failure and death (Alberta-trained KDpredict with 4 variables).

Figure S3: Distribution of study participants and estimated actual 5-year risks of kidney failure and death by CKD stage, albuminuria (external cohorts)



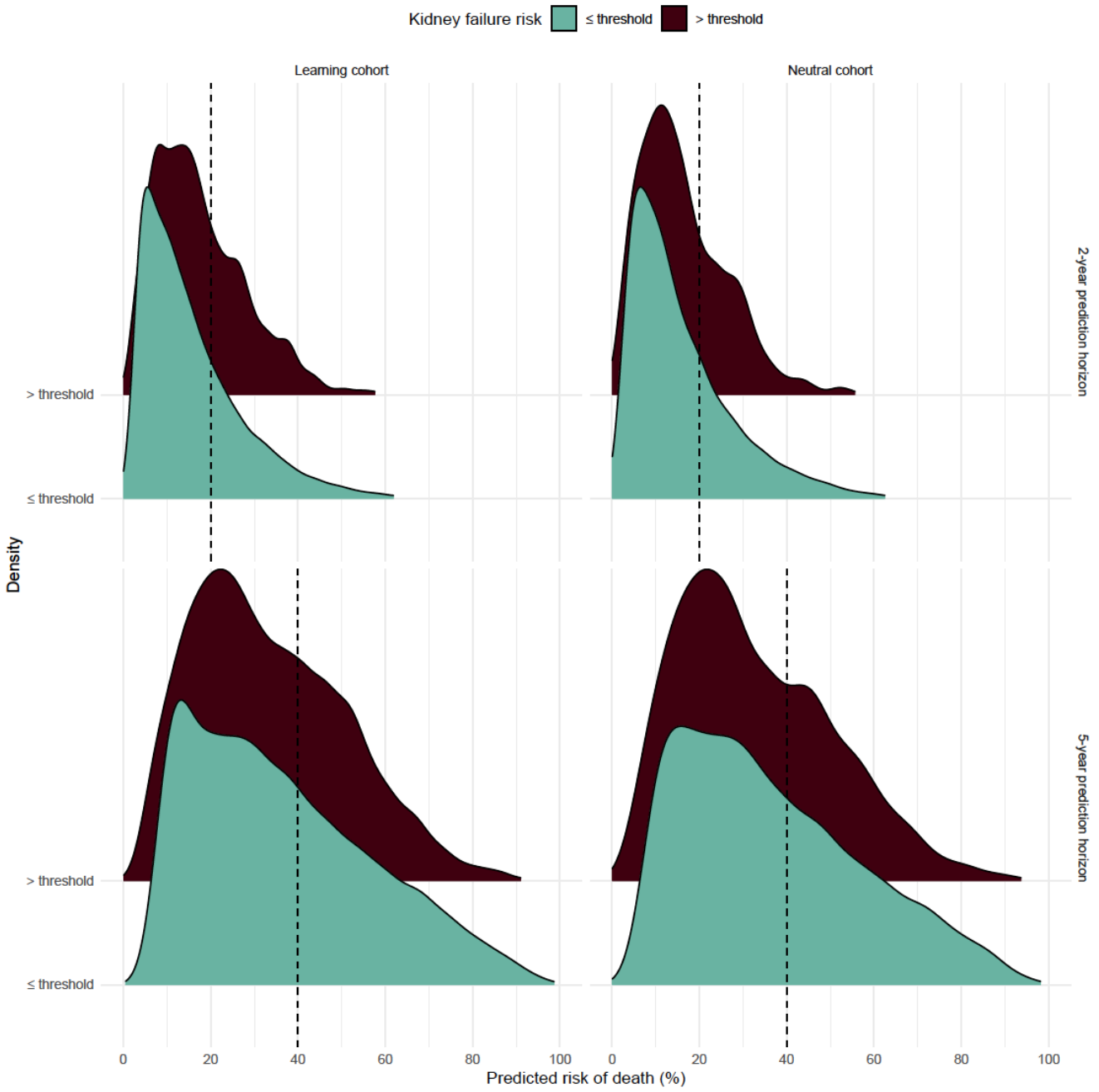
Legend: Data are from Denmark (left; N=17,528) and Scotland (right; N=7,740). G3b indicates moderate chronic kidney disease (eGFR 30-44 mL/min/1.73 m^2); G4 indicates severe chronic kidney disease (eGFR 15-29 mL/min/1.73 m^2); ACR indicates albumin-to-creatinine ratio (A1 <30 mg/g, A2 30-300 mg/g, A3 >300 mg/g). Absolute frequencies (top panels) refer to number of people; percentages (bottom panels) refer to 5-year estimated actual risks of kidney failure and death.

Figure S4: Scatter plots of predicted risks of kidney failure and death at 2 and 5 years (Alberta cohorts)



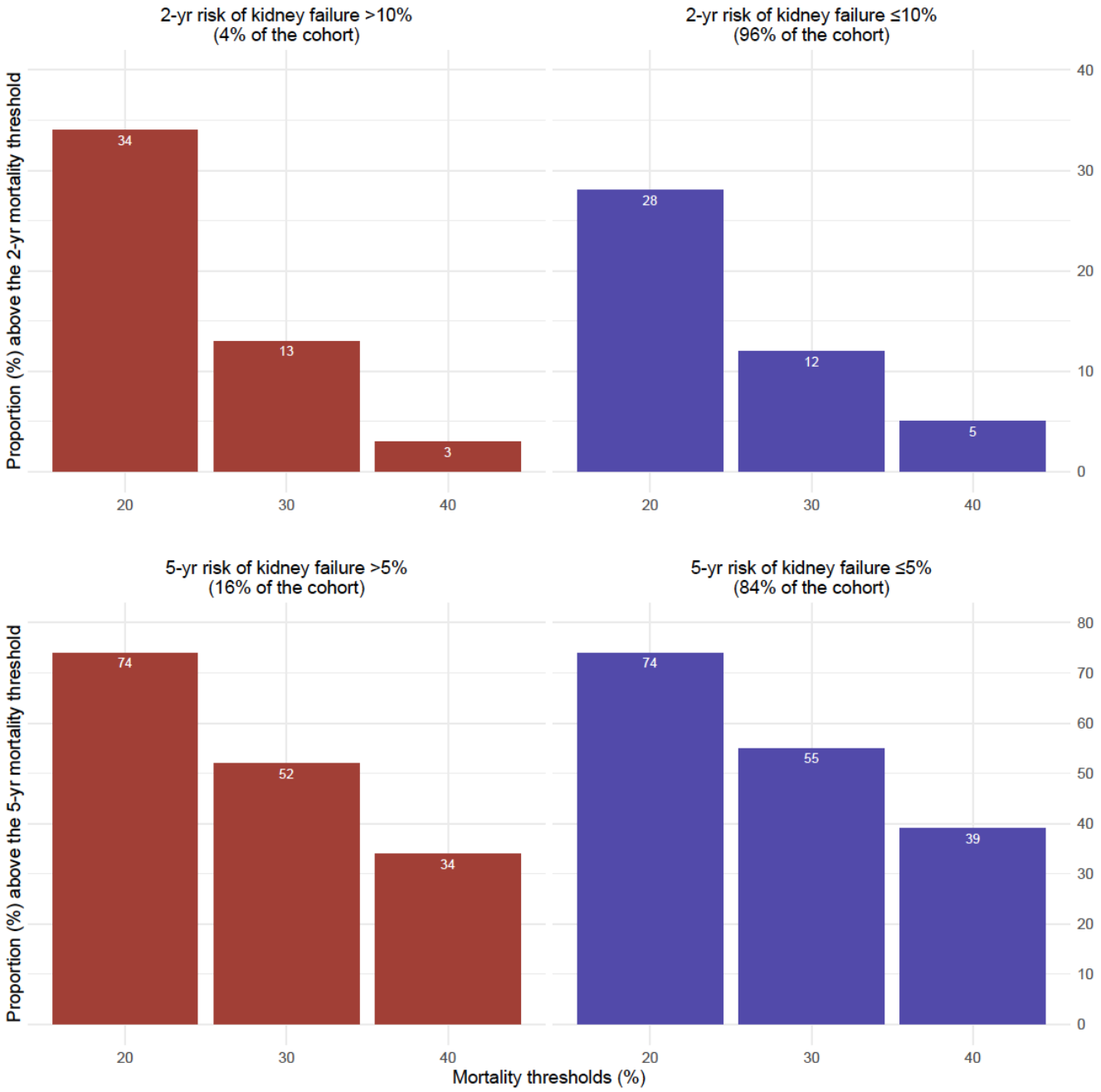
Legend: Data are from Alberta cohort data used for supervised learning (left panels) and to generate synthetic data for unsupervised learning (neutral data that the model has not seen, right panels). G3b indicates moderate chronic kidney disease (eGFR 30-44 ml/min/1.73 m^2); G4 indicates severe chronic kidney disease (eGFR 15-29 ml/min/1.73 m^2); ACR indicates albumin-to-creatinine ratio (A1 <30 mg/g, A2 30-300 mg/g, A3 >300 mg/g). Predictions were obtained from the 4-variable super-learner trained using the supervised learning cohort. Vertical dashed lines indicate current kidney failure risk thresholds, 10% at 2 years for referral to multidisciplinary clinic and preparation for management of kidney failure and 5% at 5 years for referral to nephrology care from general practice. Horizontal dashed lines indicate proposed mortality thresholds, 20% at 2 years and 40% at 5 years.

Figure S5: Distribution of predicted mortality by predicted risk of kidney failure



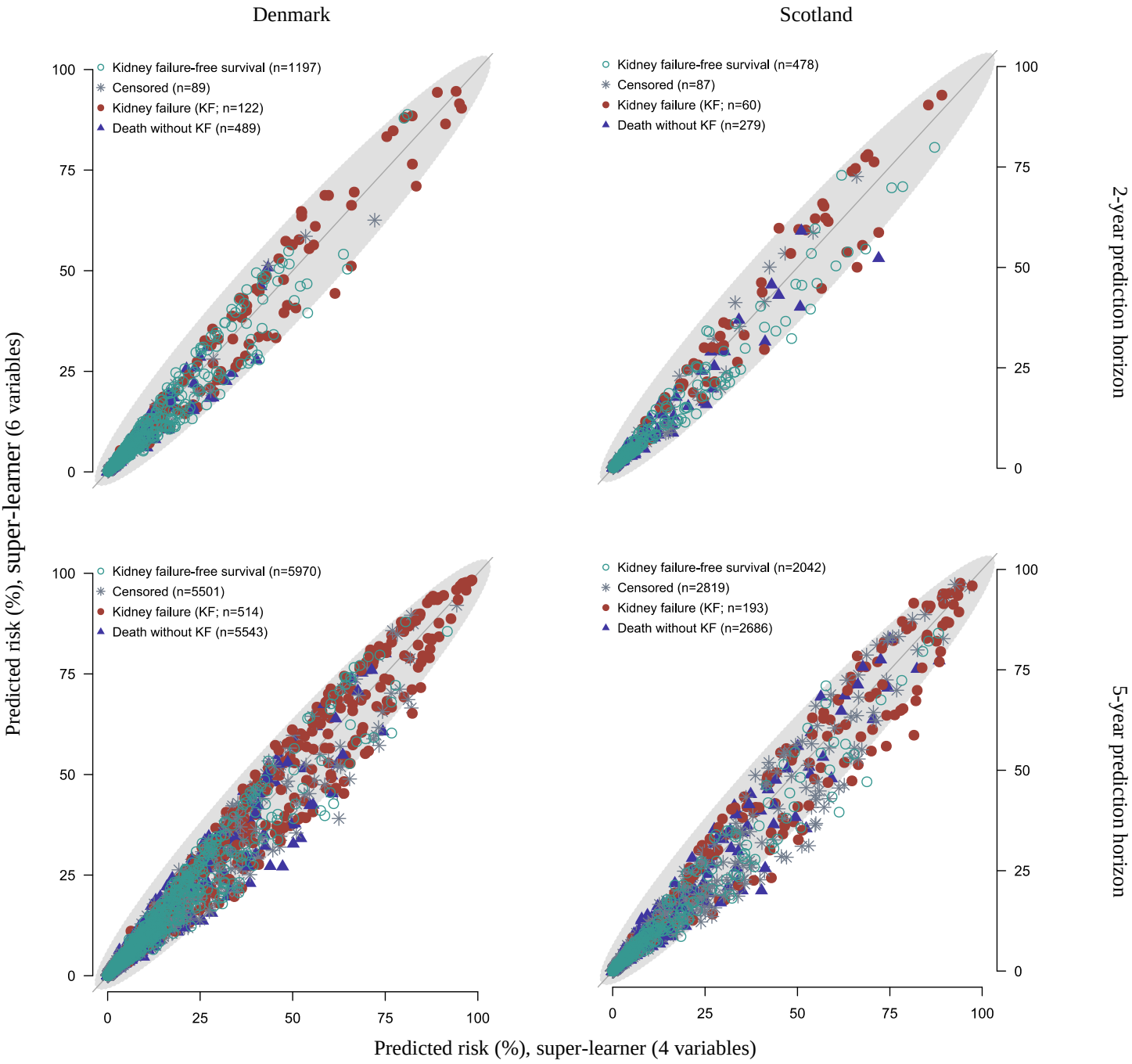
Legend: Data are from Alberta cohort data used for supervised learning (left panels) and to generate synthetic data for unsupervised learning (neutral data that the model has not seen, right panels). Kidney failure risk thresholds are 10% at 2 years (referral to enhanced multidisciplinary clinic) and 5% at 5 years (referral to nephrology from general practice). Vertical dashed lines indicate proposed mortality thresholds, 20% at 2 years and 40% at 5 years. Predictions are from the 4-variable KDpredict applied to the Alberta cohort.

Figure S6: Proportion of people above proposed mortality thresholds according to the risk of kidney failure



Legend: Bars represent the proportions (%) of people above different proposed thresholds for predicted risk of death (arbitrary values of 20%, 30%, 40%), among those who did (left) and did not meet (right) referral thresholds for multidisciplinary care (10% at 2 years) or nephrology care (5% at 5 years). Predictions are from the 4-variable KDpredict applied to the Alberta cohort.

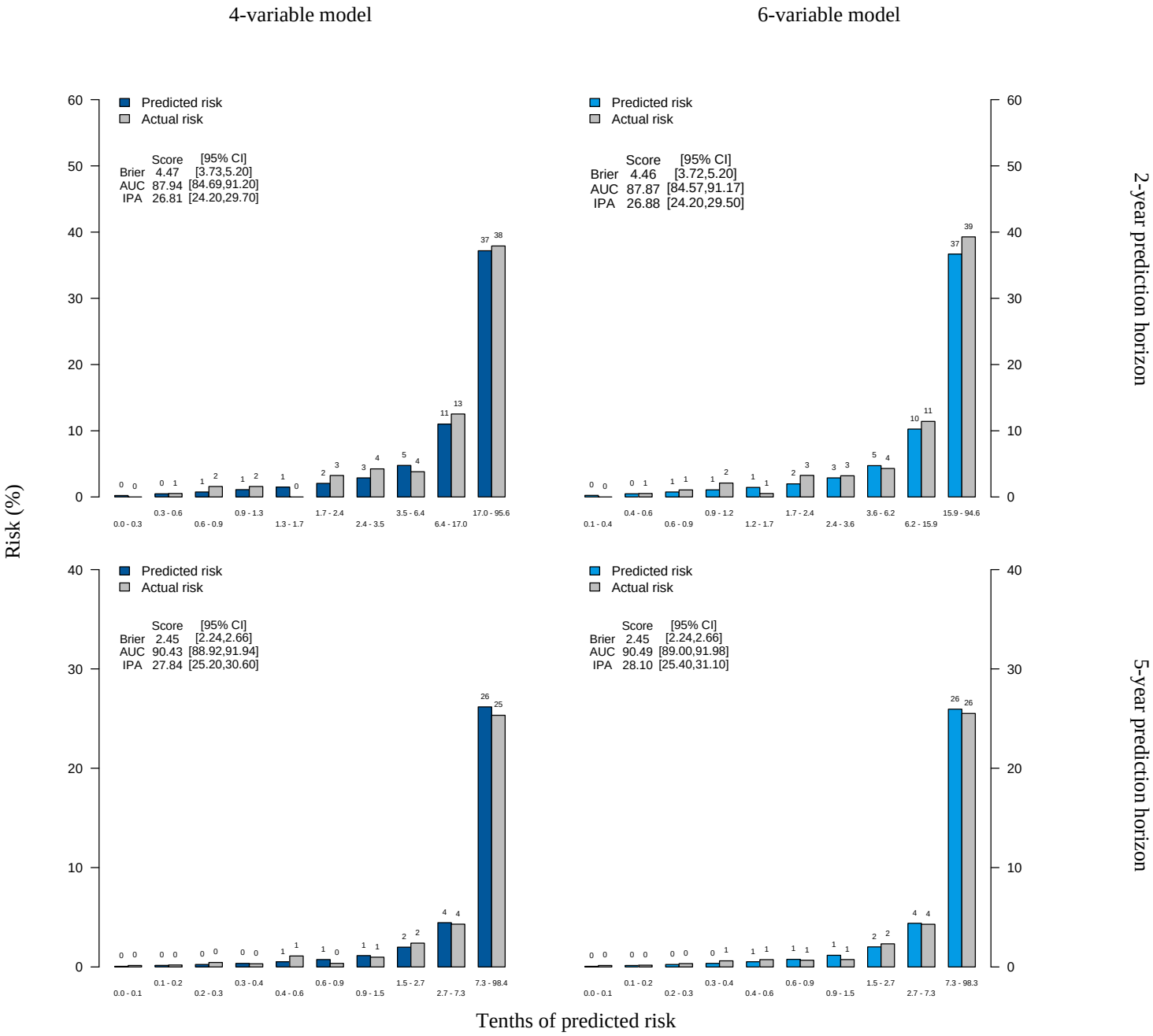
Figure S7: Agreement between individual 2- and 5-year risk predictions of kidney failure (4- vs 6-variable super-learner)



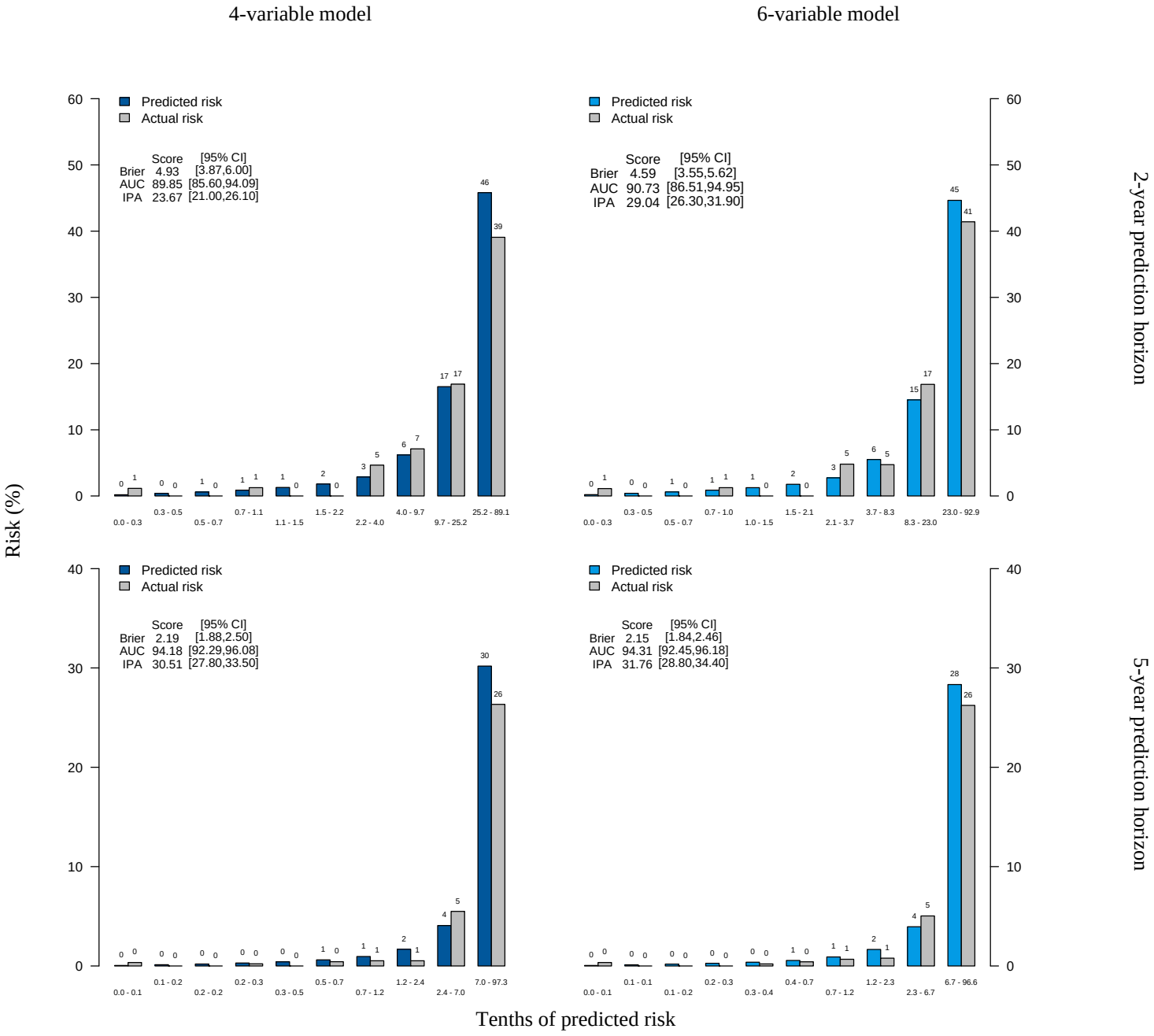
Legend: Scatterplot of predicted risks of kidney failure by site (Denmark cohort, left panels, and Scotland cohort, right panels) and prediction time horizon (2 years for G4 CKD, top, and 5 years for the overall G3bG4 cohort, bottom). Each point indicates the status of an individual at the prediction time horizon: alive without kidney failure (kidney failure-free survival), censored (unknown status), kidney failure and competing risk (death without kidney failure). The gray-shaded region in each plot indicates the area of clinically meaningless risk difference (set at the pre-specified value of 10%). The models were trained in Alberta, Canada, and applied in Denmark and Scotland.

Figure S8: Calibration of 4- and 6-variable super-learner for 2- and 5-year prediction of kidney failure

A. Denmark cohort



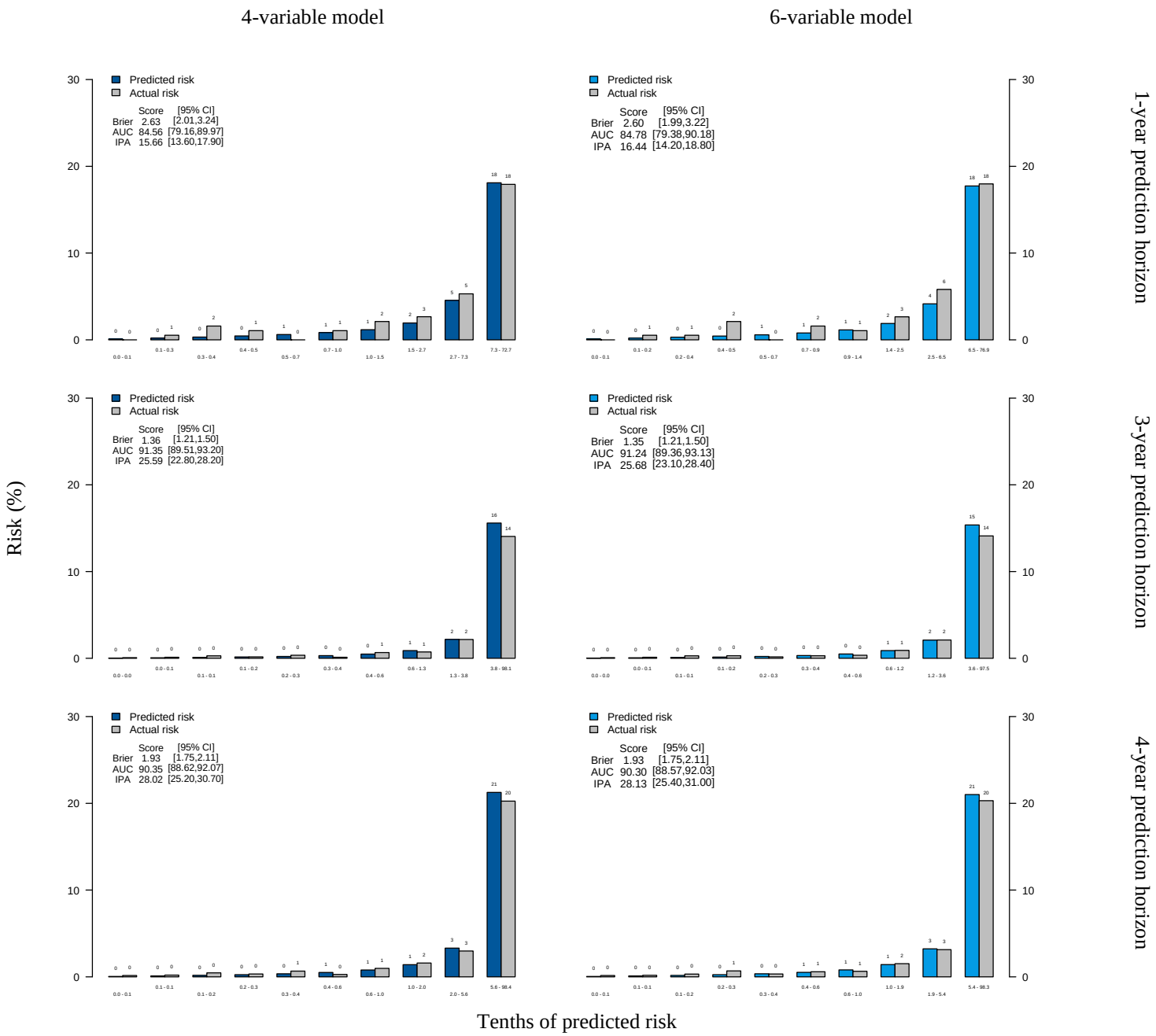
B. Scotland cohort



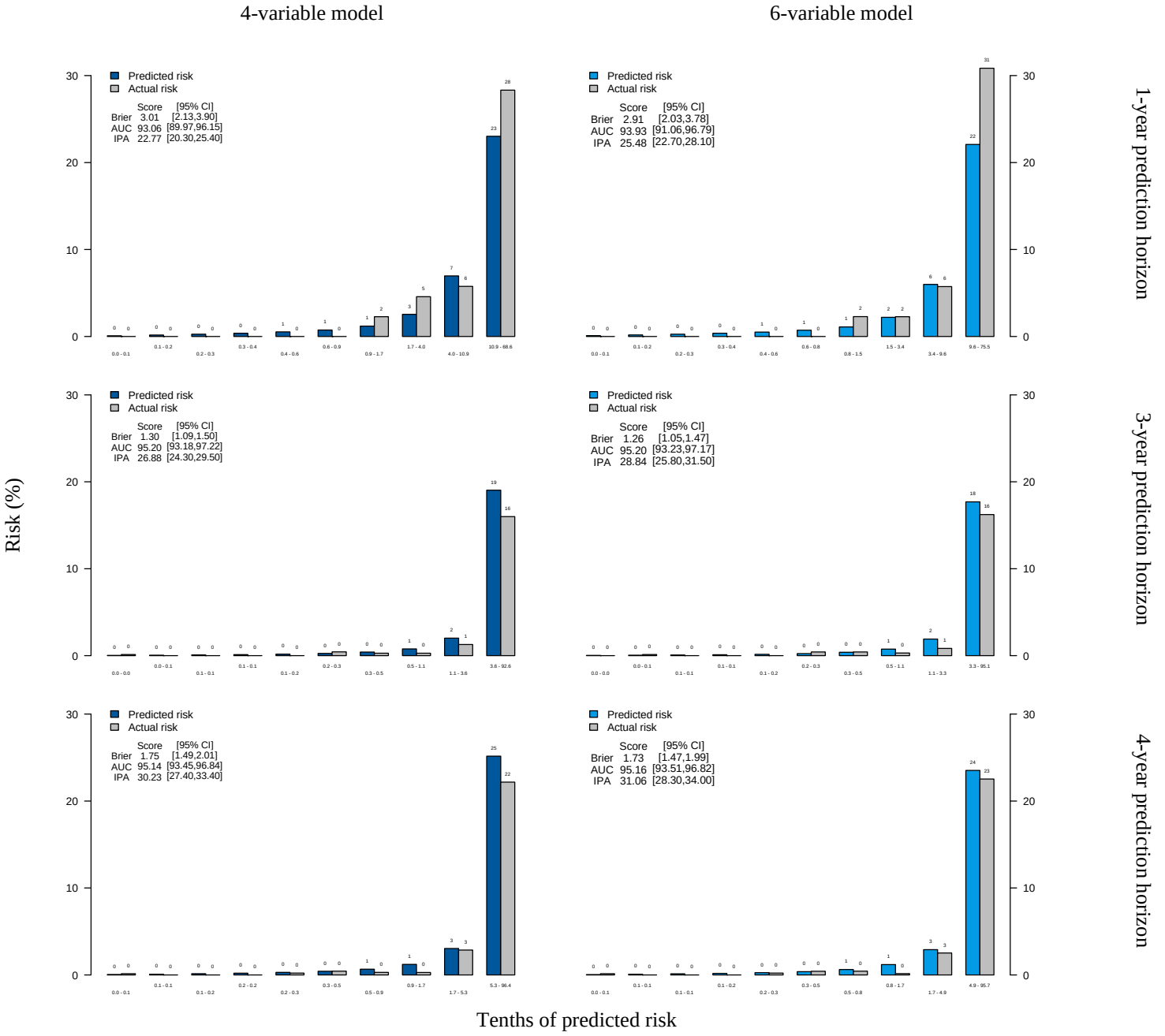
Legend: The models were trained in Alberta and tested on the full set of external data, Denmark (A) and Scotland (B). Prediction time horizons: 2 years for G4 CKD (top) and 5 years for the whole cohort (bottom). Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars)

Figure S9: Calibration of 4- and 6-variable super-learner for 1-, 3- and 4-year prediction of kidney failure

A. Denmark cohort



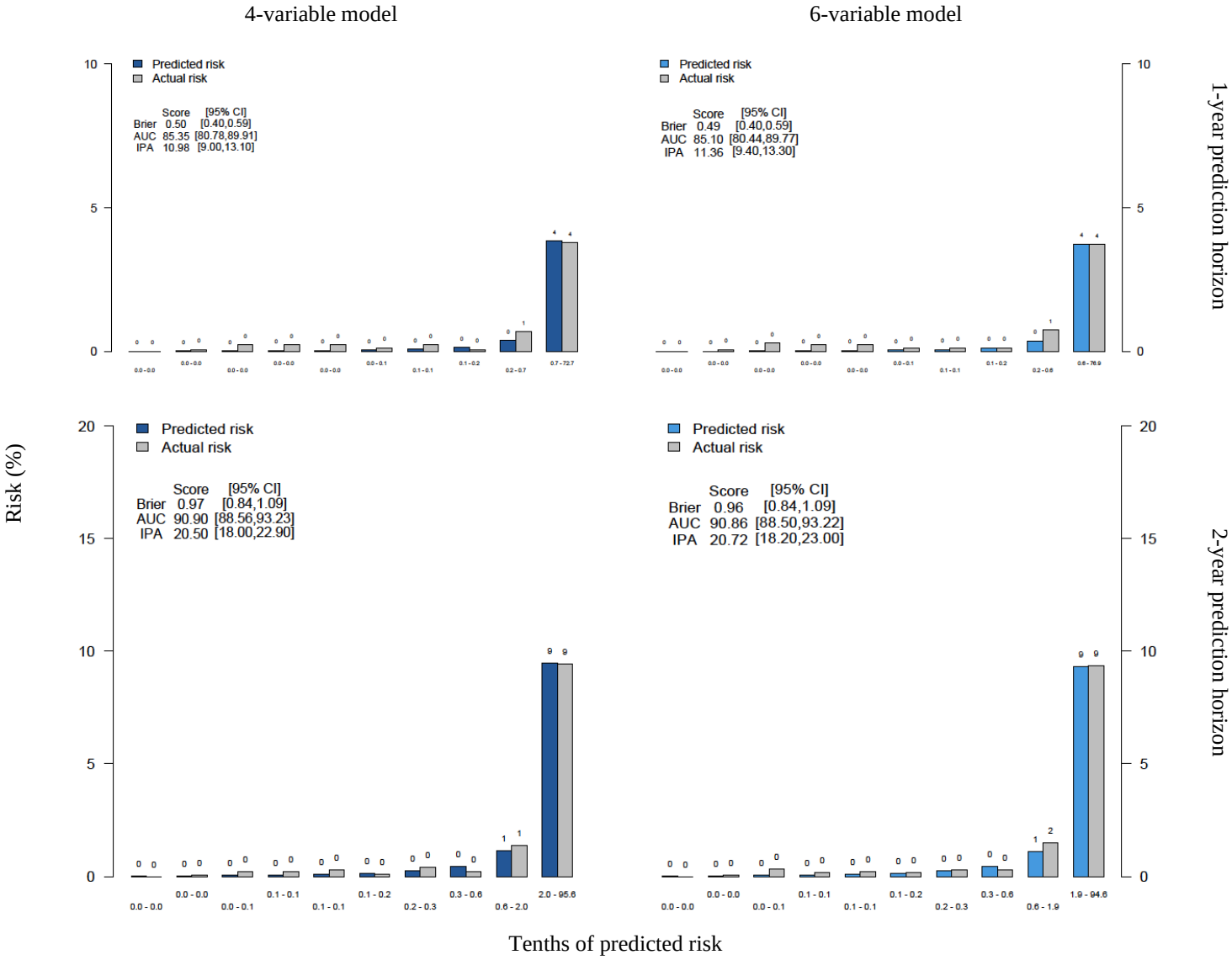
B. Scotland cohort



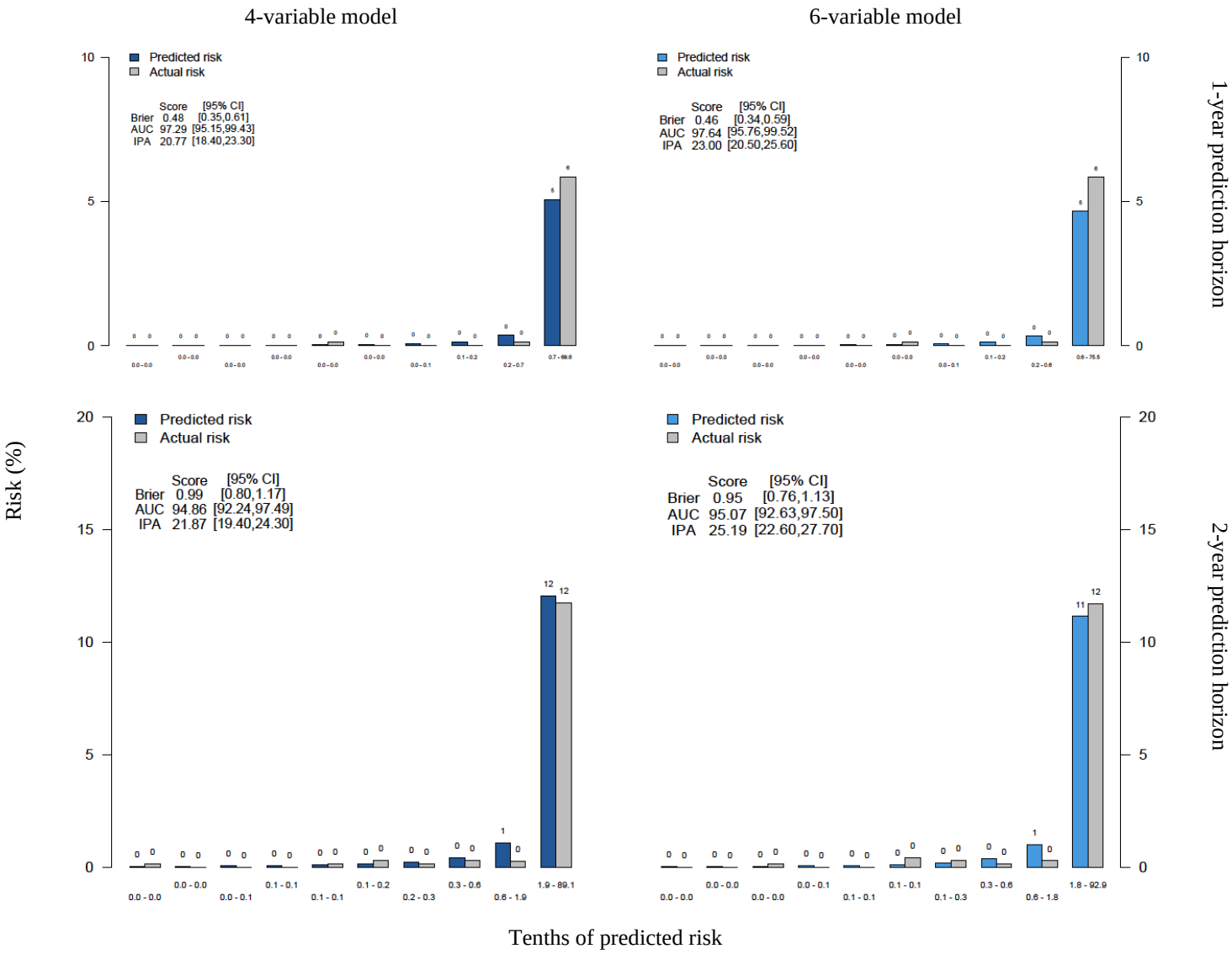
Legend: The models were trained in Alberta and tested on the full set of external data, Denmark (A) and Scotland (B). Prediction time horizons: 1 years for G4 CKD (top) and 3 and 4 years for the whole cohort (bottom). Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars).

Figure S10: Calibration of 4- and 6-variable super-learner for 1- and 2-year prediction of kidney failure in the full G3bG4-CKD cohort

A. Denmark cohort

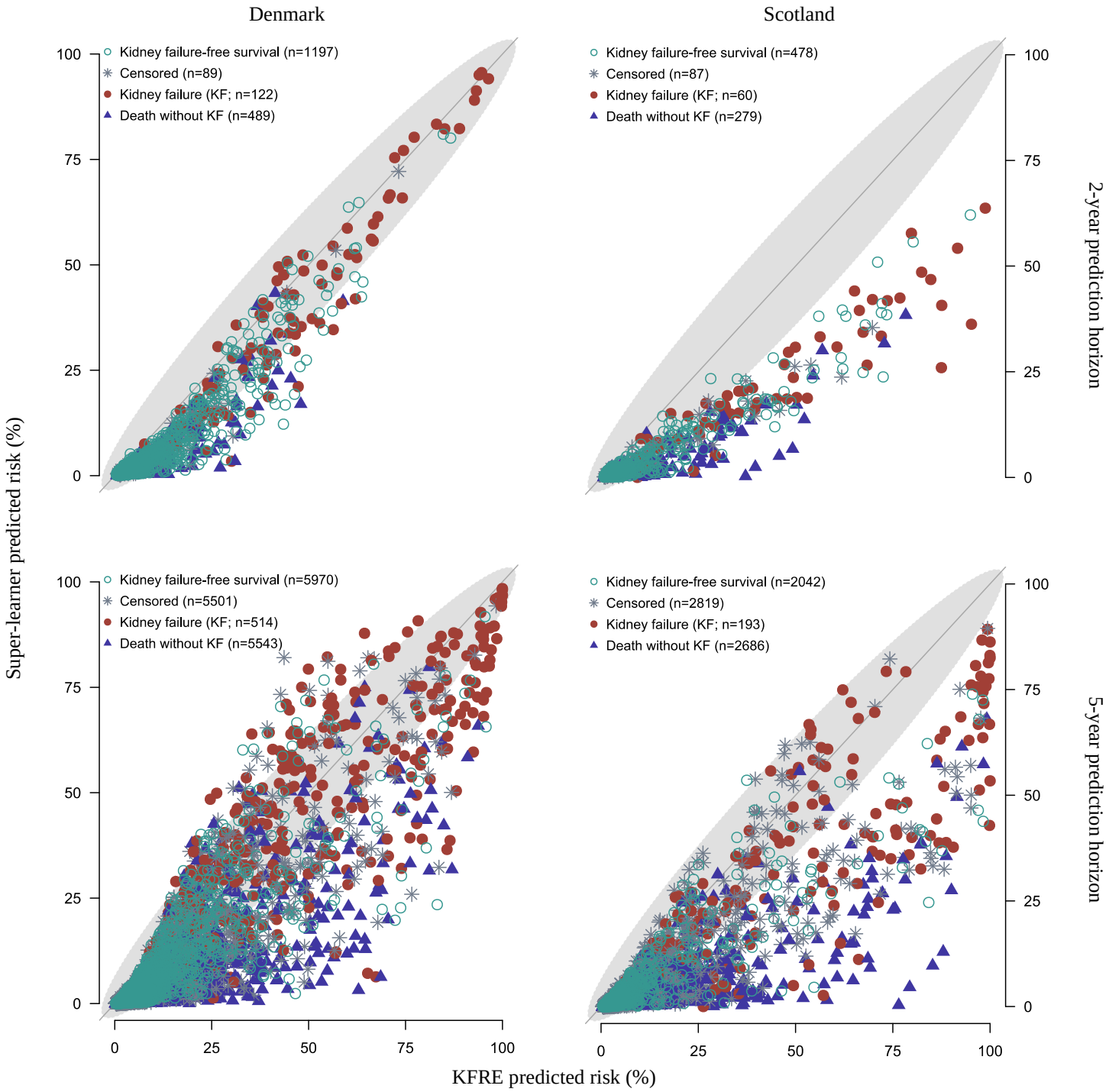


B. Scotland cohort



Legend: The models were trained in Alberta and tested on the full set of external data, Denmark (A) and Scotland (B). Prediction time horizons: 1 and 2 years for the whole cohort (G3bG4-CKD). Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars)

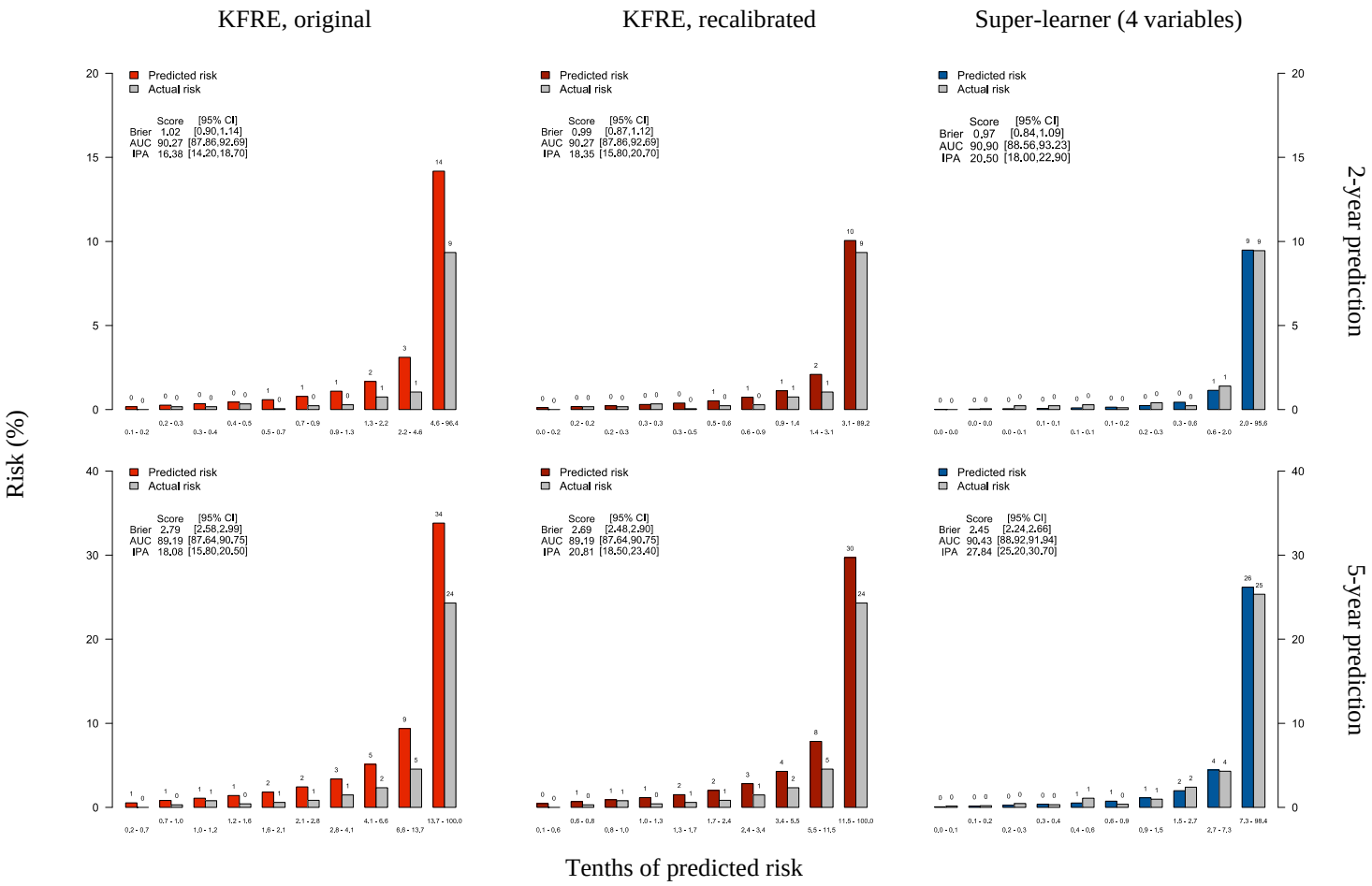
Figure S11: Agreement between individual 2- and 5-year risk predictions of kidney failure (original KFRE vs 4-variable super-learner)



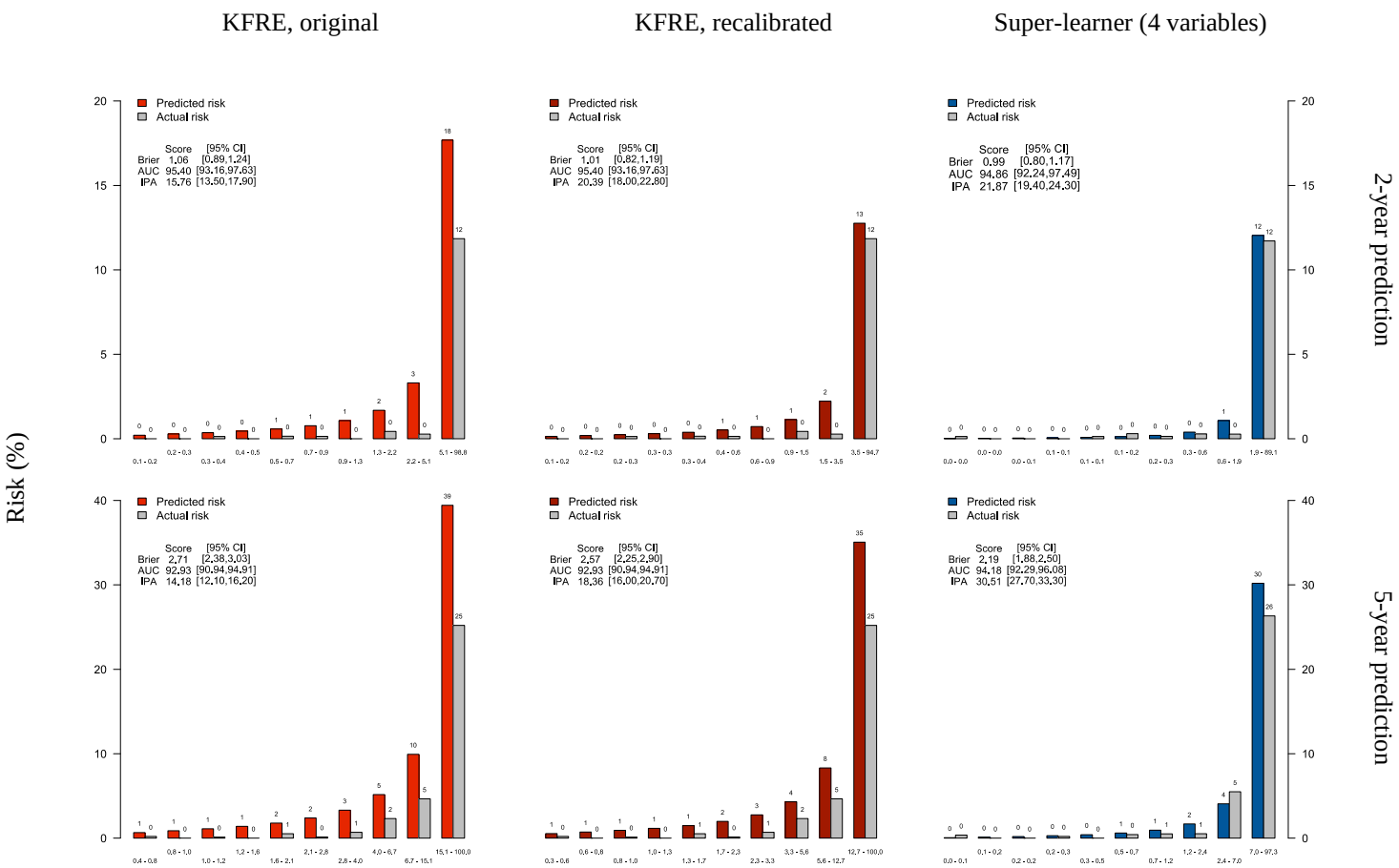
Legend: Scatterplot of predicted risks of kidney failure by site (Denmark cohort, left panels, and Scotland cohort, right panels) and prediction horizon (2 years for G4 CKD, top, and 5 years for the whole G3bG4 cohort, bottom). KFRE, original 4-variable kidney failure risk equation;[11] super-learner, 4-variable model trained in Alberta. Each point indicates the status of an individual at the prediction time: kidney failure-free survival (alive without kidney failure), censored (unknown status), kidney failure, and death without kidney failure (competing risk). The gray-shaded region is the area of clinically meaningless risk difference (set at the prespecified value of 10%).

Figure S12: Original and recalibrated KFRE vs 4-variable super-learner for 2- and 5-year prediction of kidney failure in people with G3bG4-CKD

A. Denmark

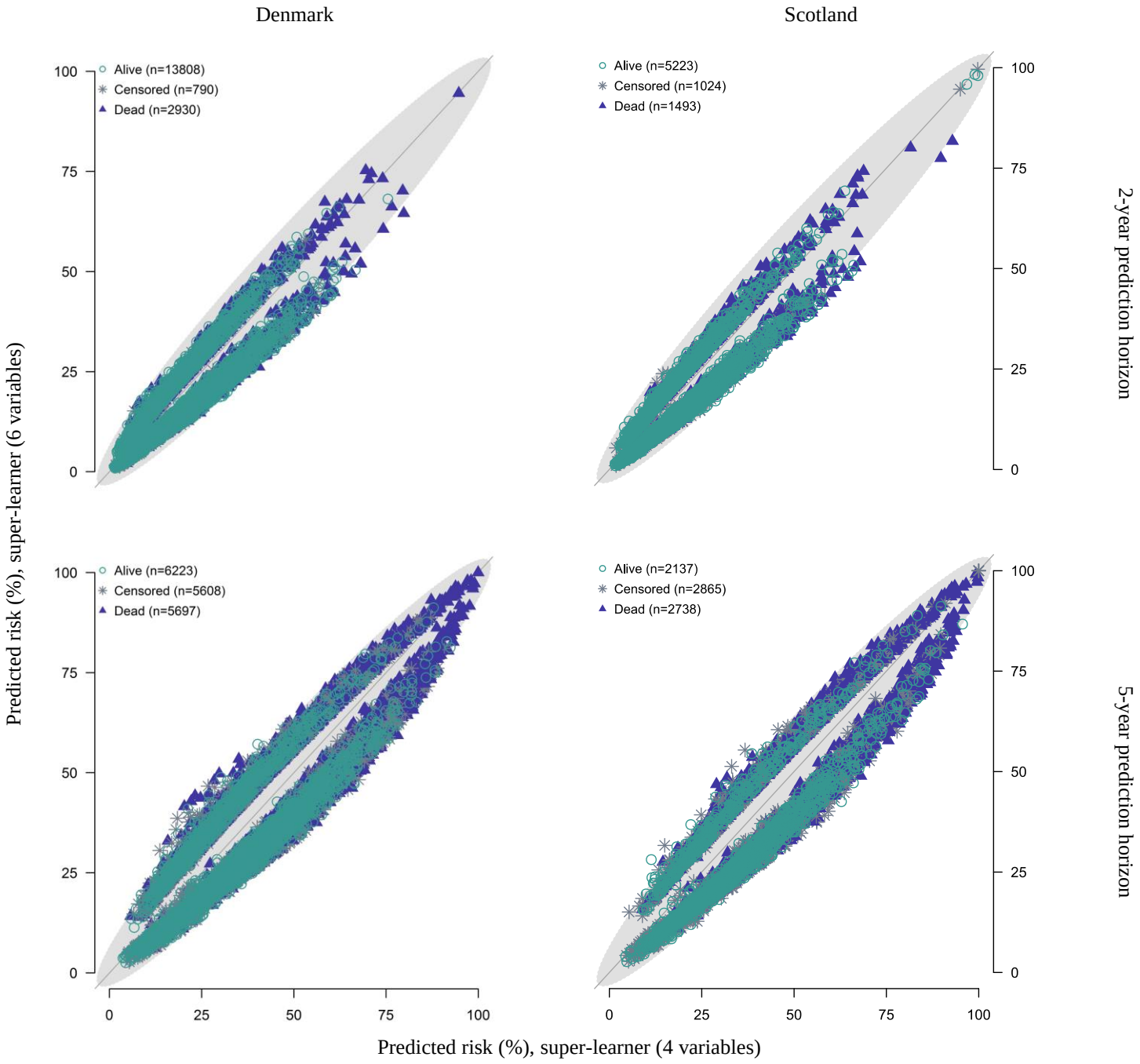


B. Scotland



Legend: KFRE indicates the 4-variable kidney failure risk equation, original equation (left panels)[11] and the same equation recalibrated for non-North-American countries (middle panels);[12] the 4-variable super-learner (right panels) was trained in Alberta, Canada and applied as is to the external data, without recalibration or retraining. The models were tested on the full set of external data, in Denmark (A) and Scotland (B) for 2- and 5-year prediction time horizons (the recalibration factor was obtained using people with G3 and G4 CKD[12]). Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars).

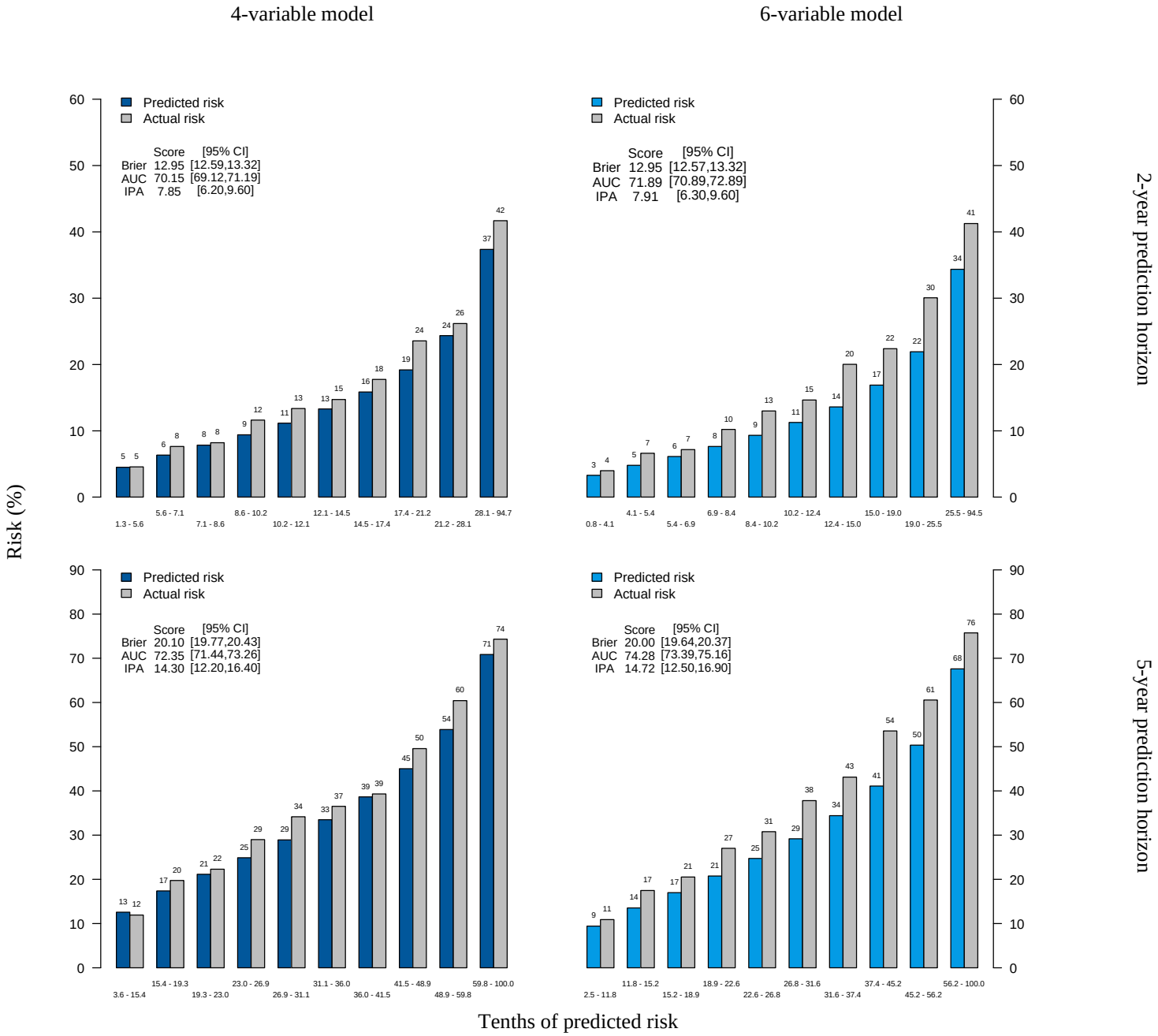
Figure S13: Agreement between individual 2- and 5-year risk predictions of death (4- vs 6-variable super-learner)



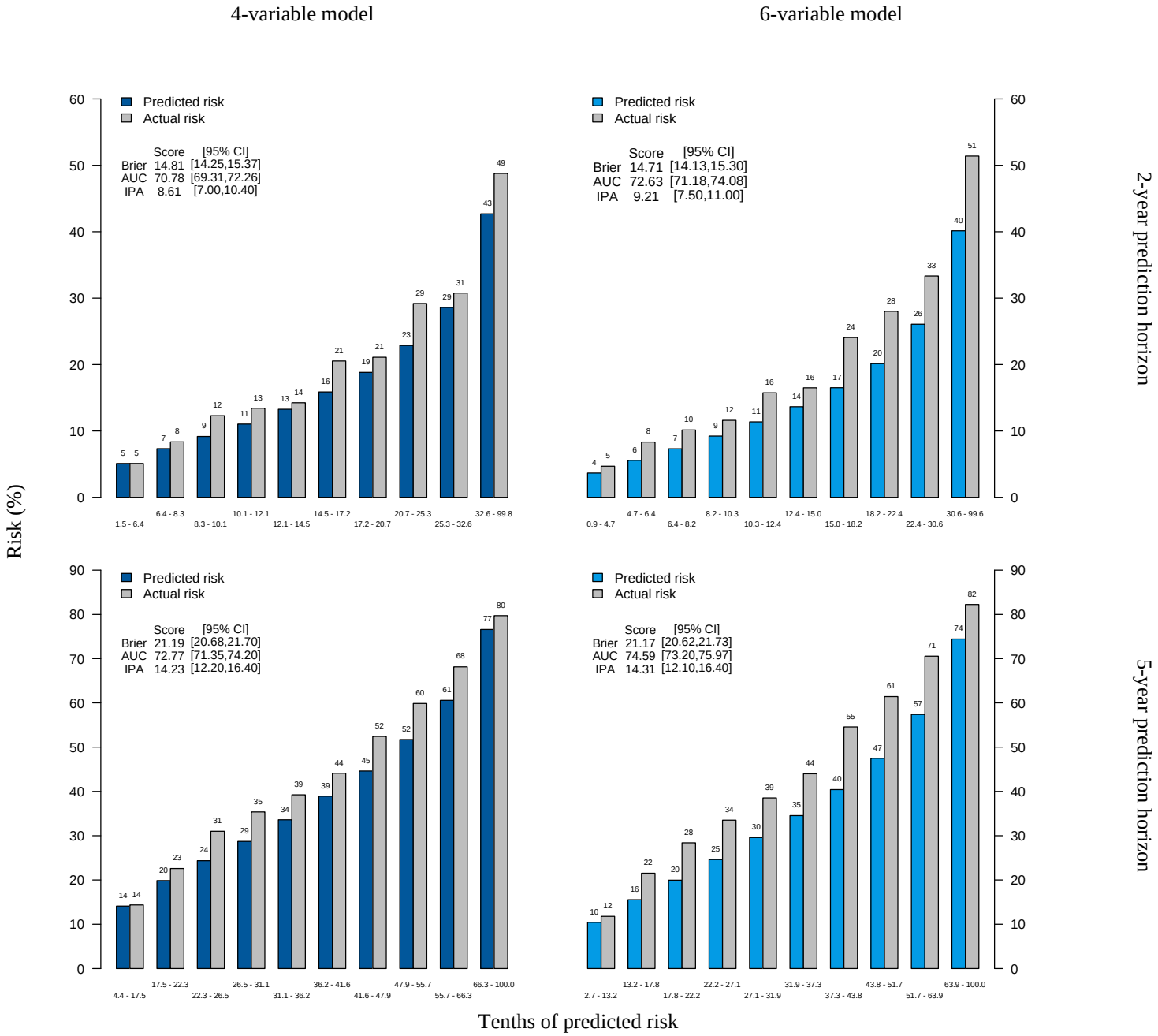
Legend: Scatterplot of predicted risks of death by site (Denmark cohort, left panels, and Scotland cohort, right panels) and prediction time horizon (2 years [top] and 5 years [bottom] for the full cohort). Each point indicates the status of an individual at the prediction time horizon: alive, censored (unknown status), or dead. The gray-shaded region in each plot indicates the area of clinically meaningless risk difference (set at the pre-specified value of 10%). The models were trained in Alberta, Canada, and applied in Denmark and Scotland.

Figure S14: Calibration of 4- vs 6-variable super-learner for 2- and 5-year prediction of death

A. Denmark cohort



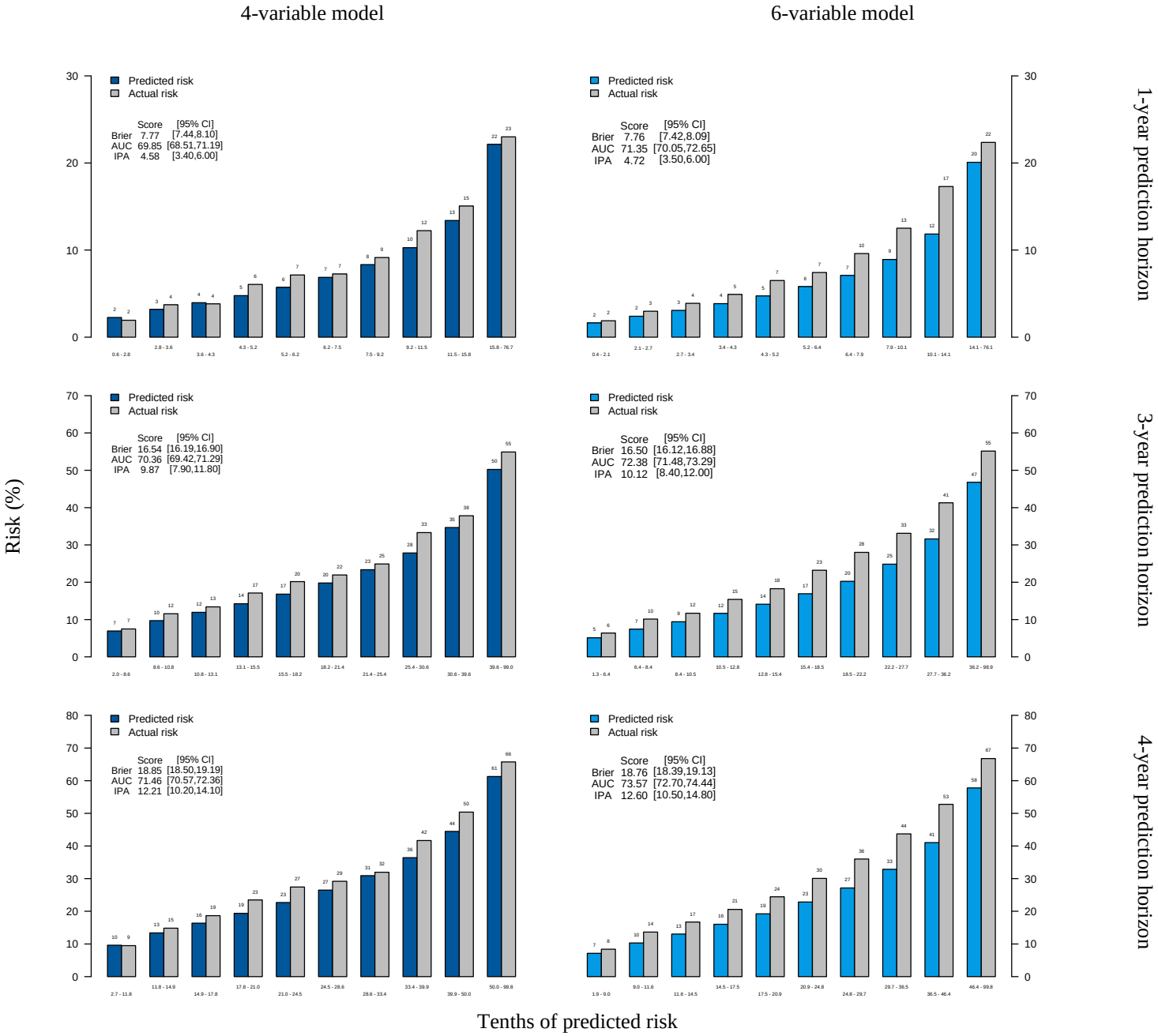
B. Scotland cohort



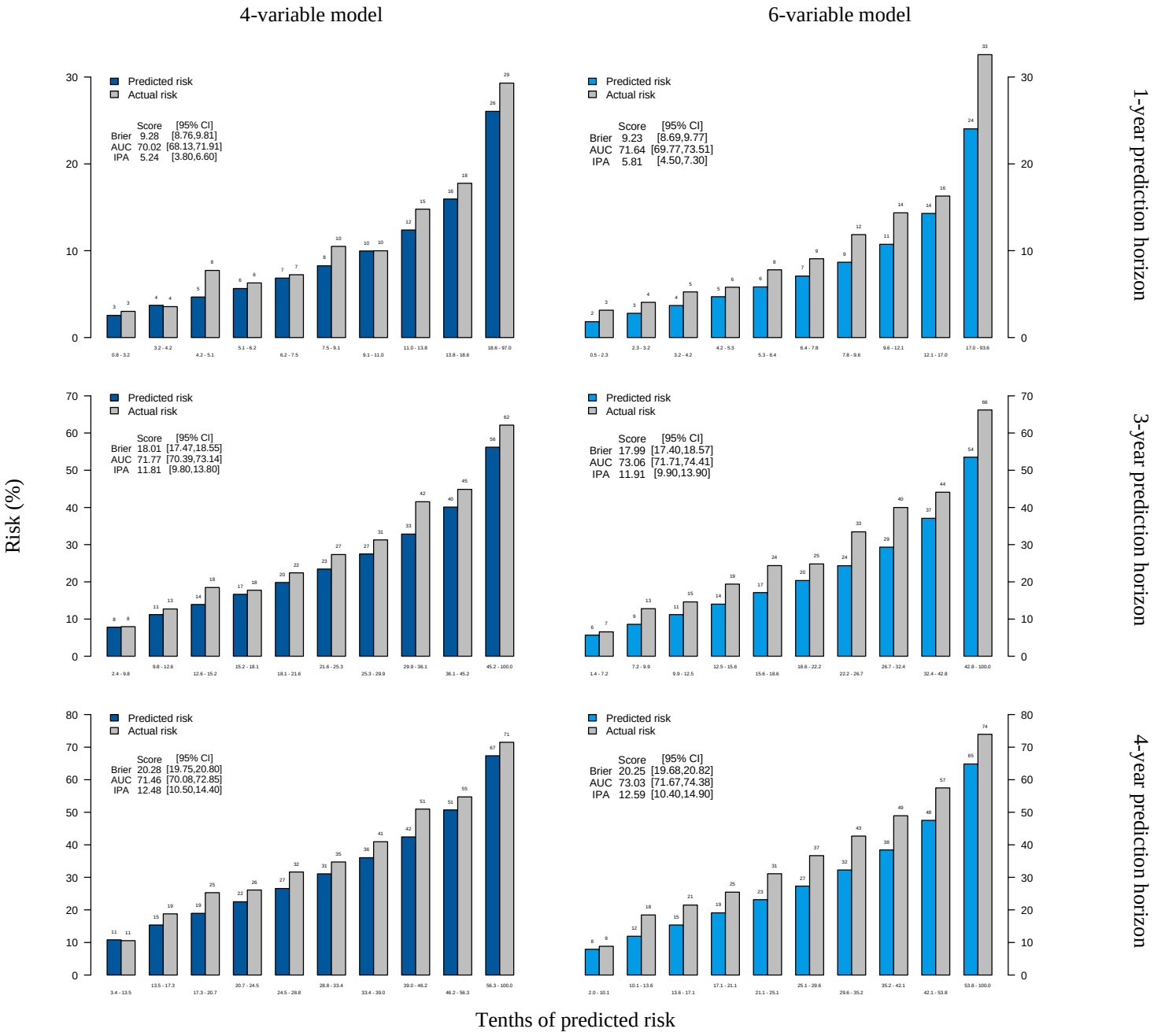
Legend: The models were trained in Alberta and tested on the full set of external data, in Denmark (A) and Scotland (B). Prediction time horizons: 2 (top) and 5 years (bottom) for the whole cohort. Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars).

Figure S15: Calibration of 4- and 6-variable super-learner for 1-, 3- and 4-year prediction of death

A. Denmark cohort

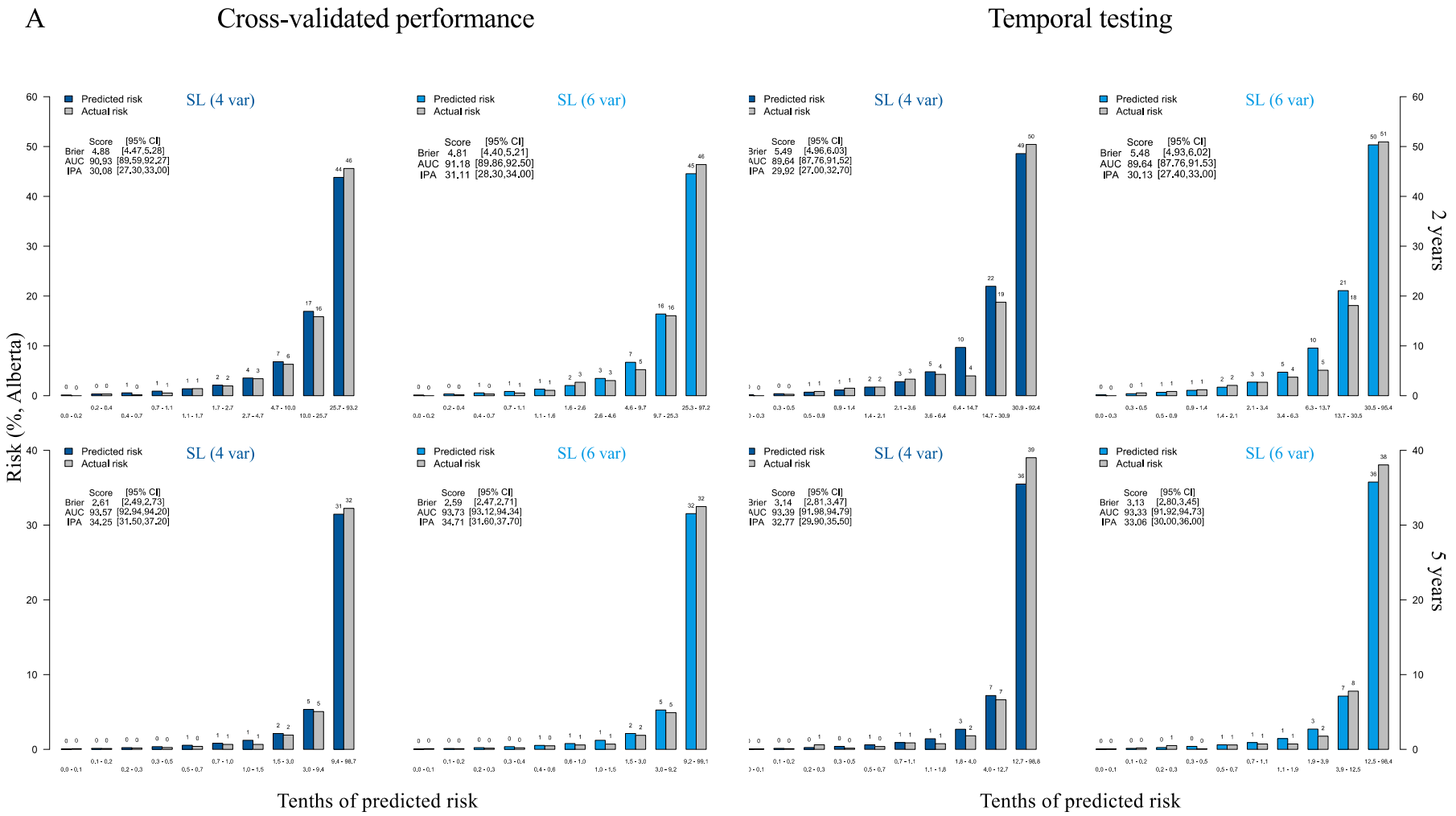


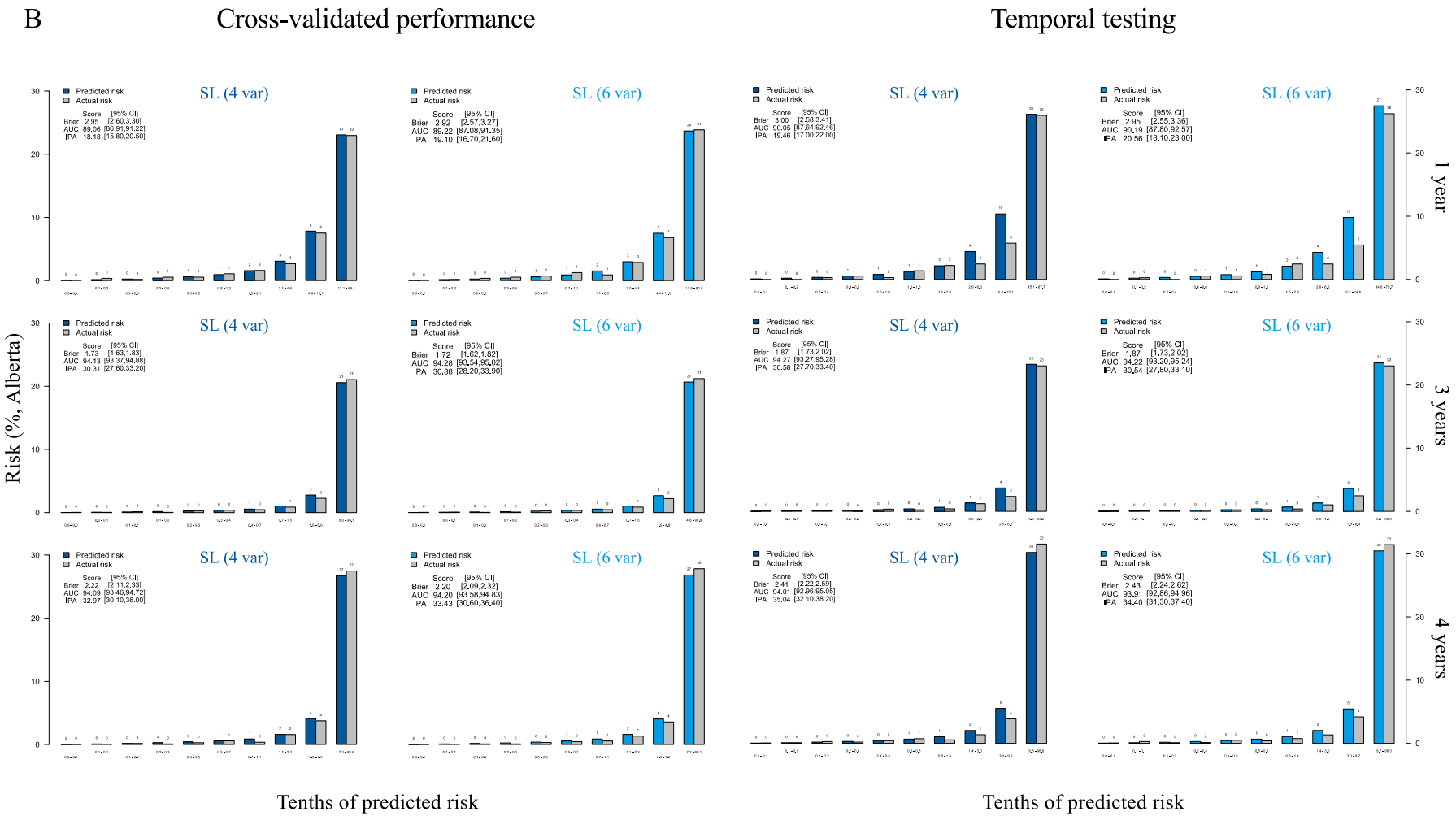
B. Scotland cohort



Legend: The models were trained in Alberta and tested on the full set of external data, in Denmark (A) and Scotland (B). Prediction time horizons: 1, 3 and 4 years for the whole cohort. Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars).

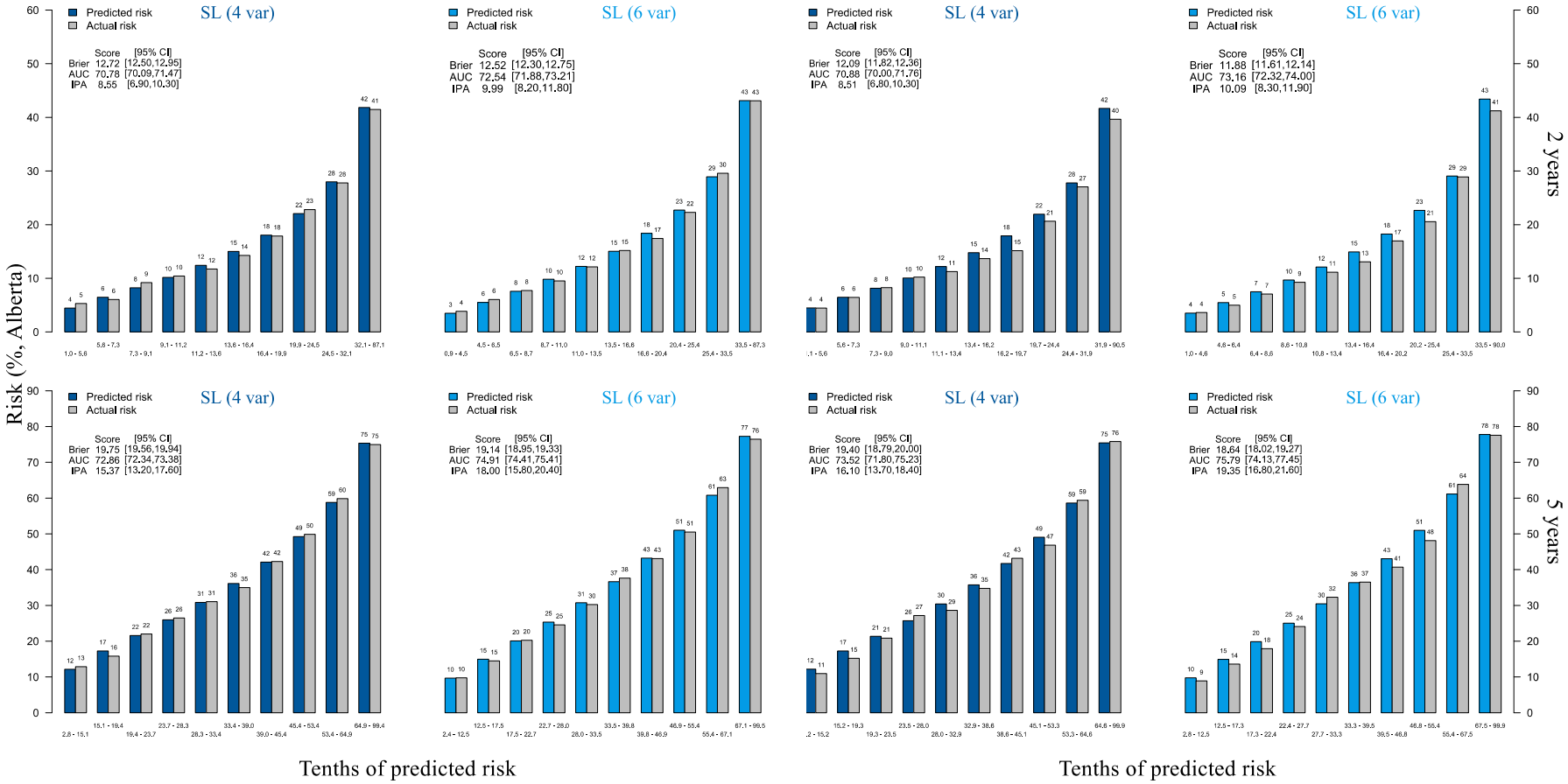
Figure S16: Temporal testing (kidney failure)

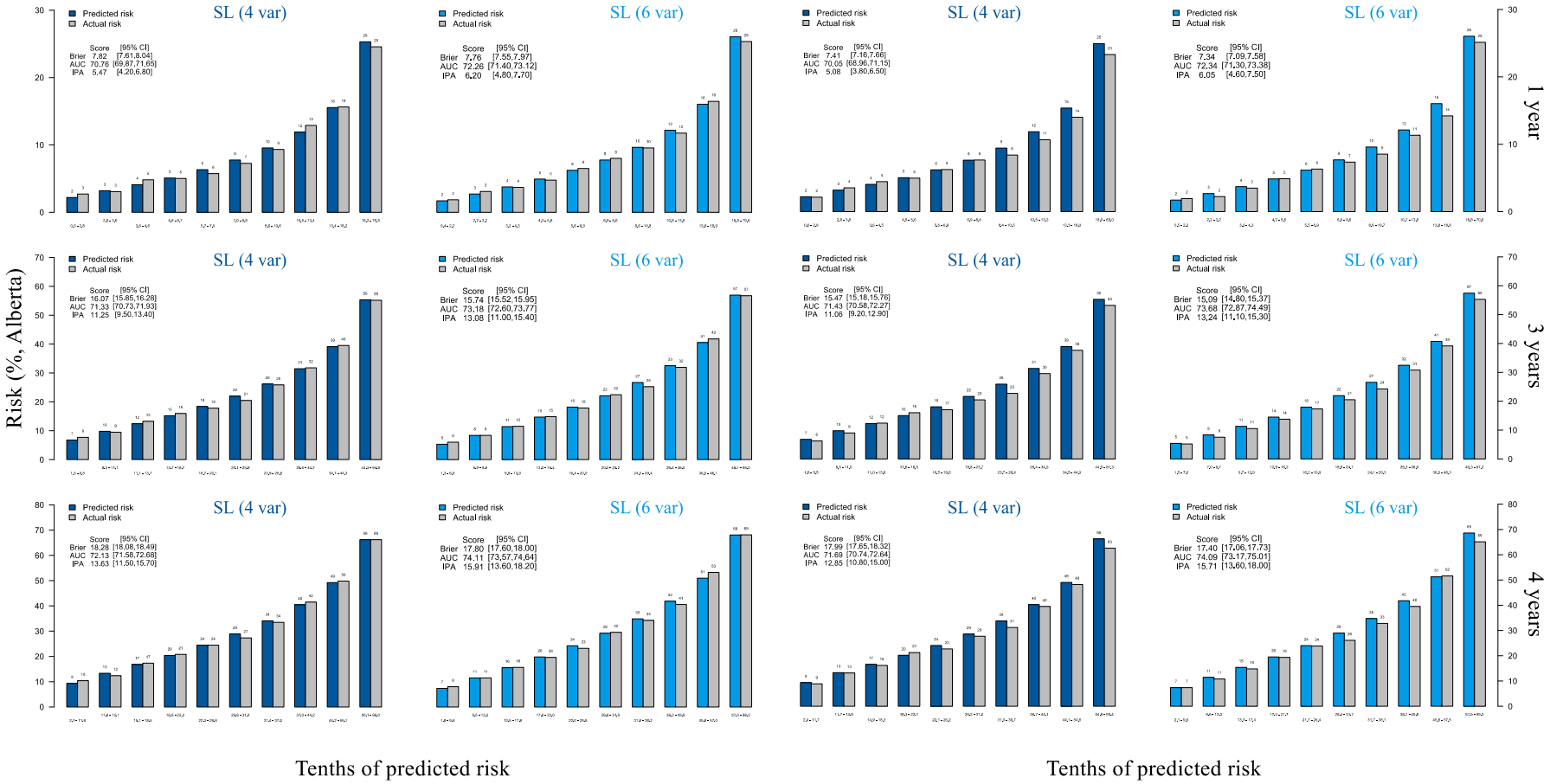




Legend: SL (4 var) indicates the 4-variable super-learner; SL (6 var) indicates the 6-variable super-learner. Prediction horizons 2 and 5 years (A) and 1, 3, and 4 years (B). Short-term (1 and 2 years) were restricted to people with G4 CKD. Left panels: cross-validated performance of the KDpredict model retrained using older Alberta data (index date until 2014-12-31). Right panels: calibration of the super-learner models retrained using older Alberta data (index date until 2014-12-31) on the full set of temporally distinct, more recent data (index date on or after 2015-01-01). Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars).

Figure S17: Temporal testing (mortality)

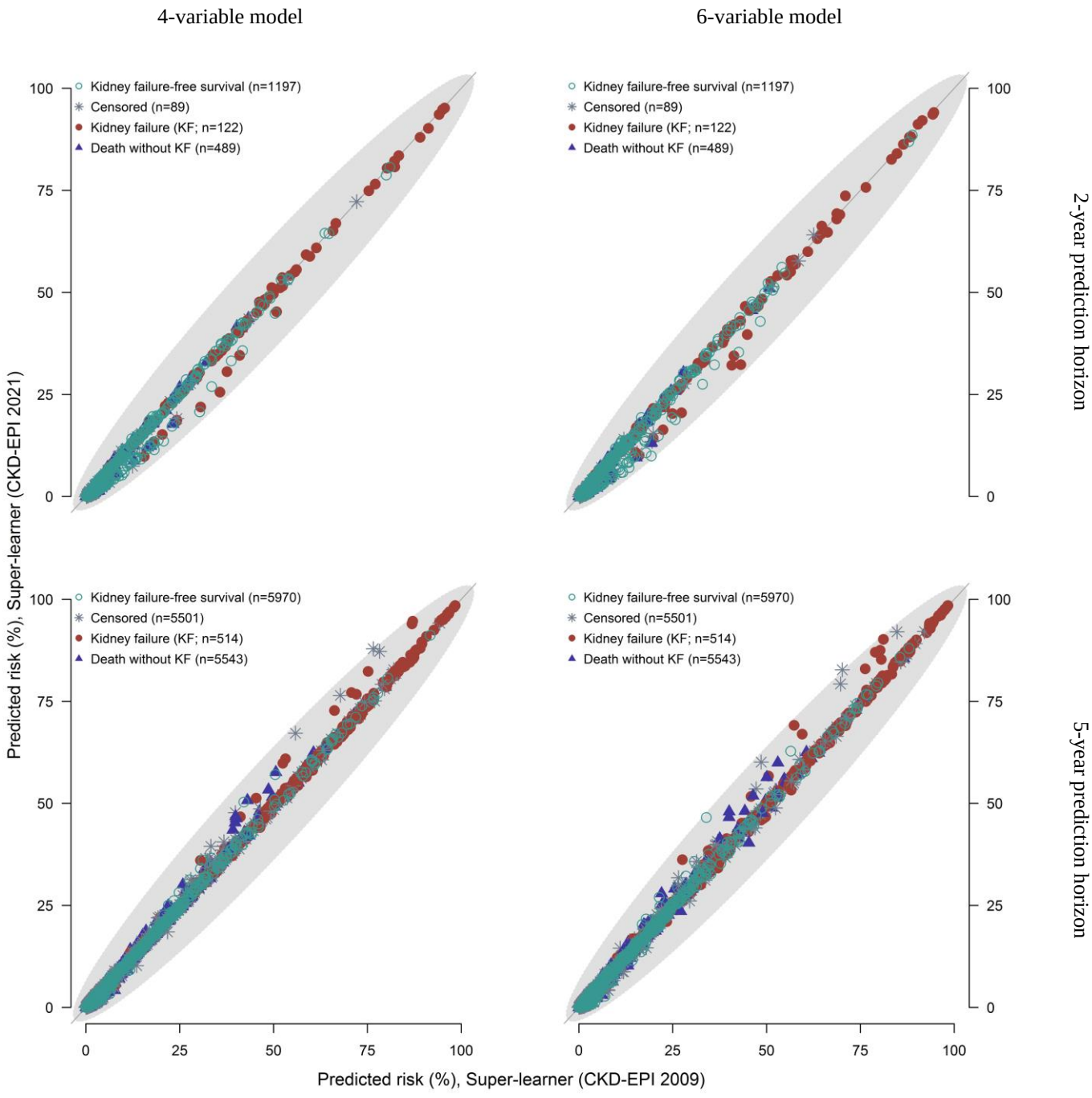




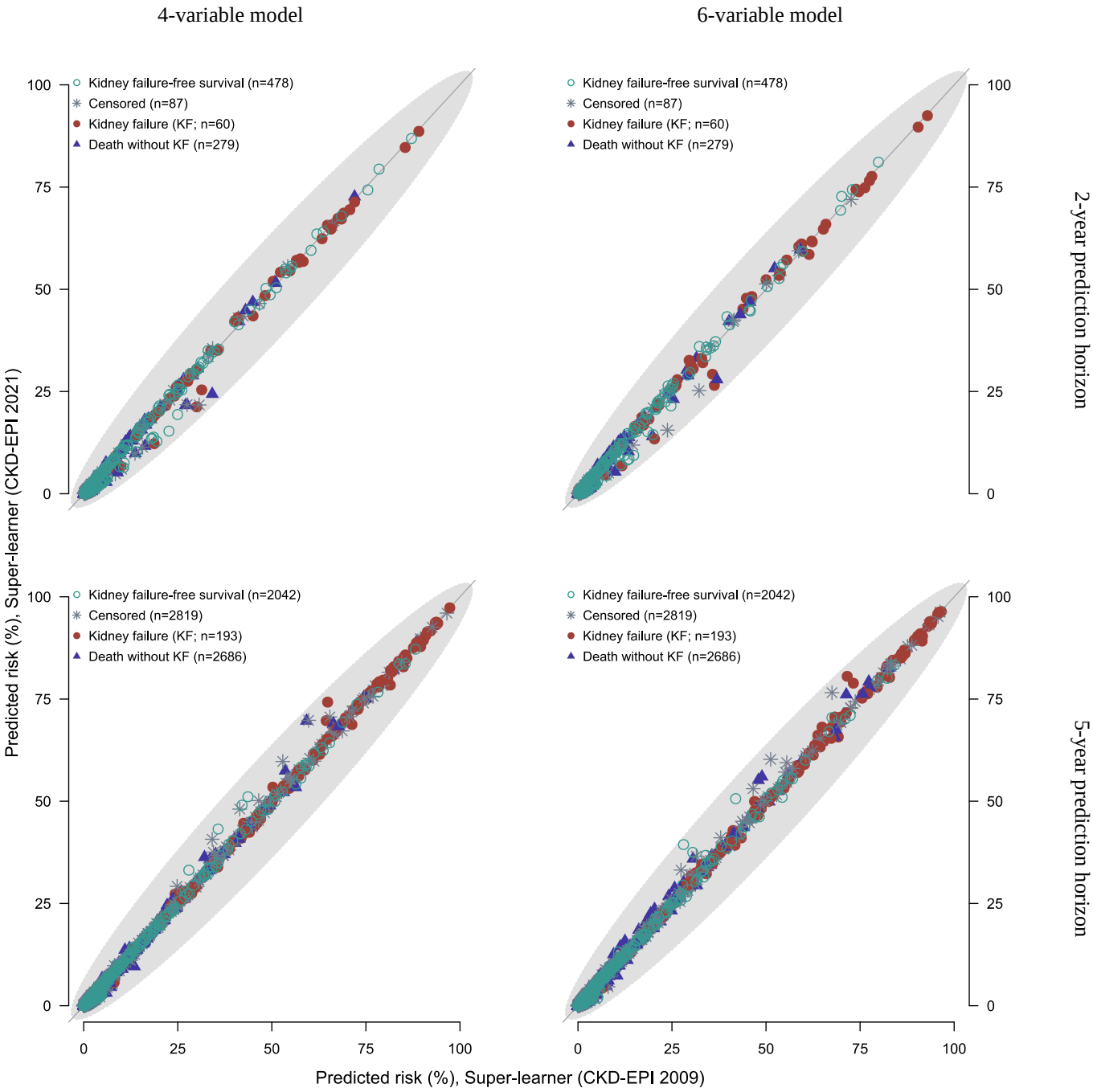
Legend: SL (4 var) indicates the 4-variable super-learner; SL (6 var) indicates the 6-variable super-learner. Prediction horizons 2 and 5 years (A) and 1, 3, and 4 years (B). Left panels: cross-validated performance of the KDpredict model retrained using older Alberta data (index date until 2014-12-31). Right panels: calibration of the super-learner models retrained using older Alberta data (index date until 2014-12-31) on the full set of temporally distinct, more recent data (index date on or after 2015-01-01). Risk predictions are grouped into 10 equally large groups (the values below the x-axis show the thresholds). Within each group, the observed frequency corresponds to the estimated actual risk (gray bars).

Figure S18: Agreement between individual 2- and 5-year risk predictions of kidney failure (eGFR calculated with EPI 2009 vs EPI 2021 formula)

A. Denmark cohort



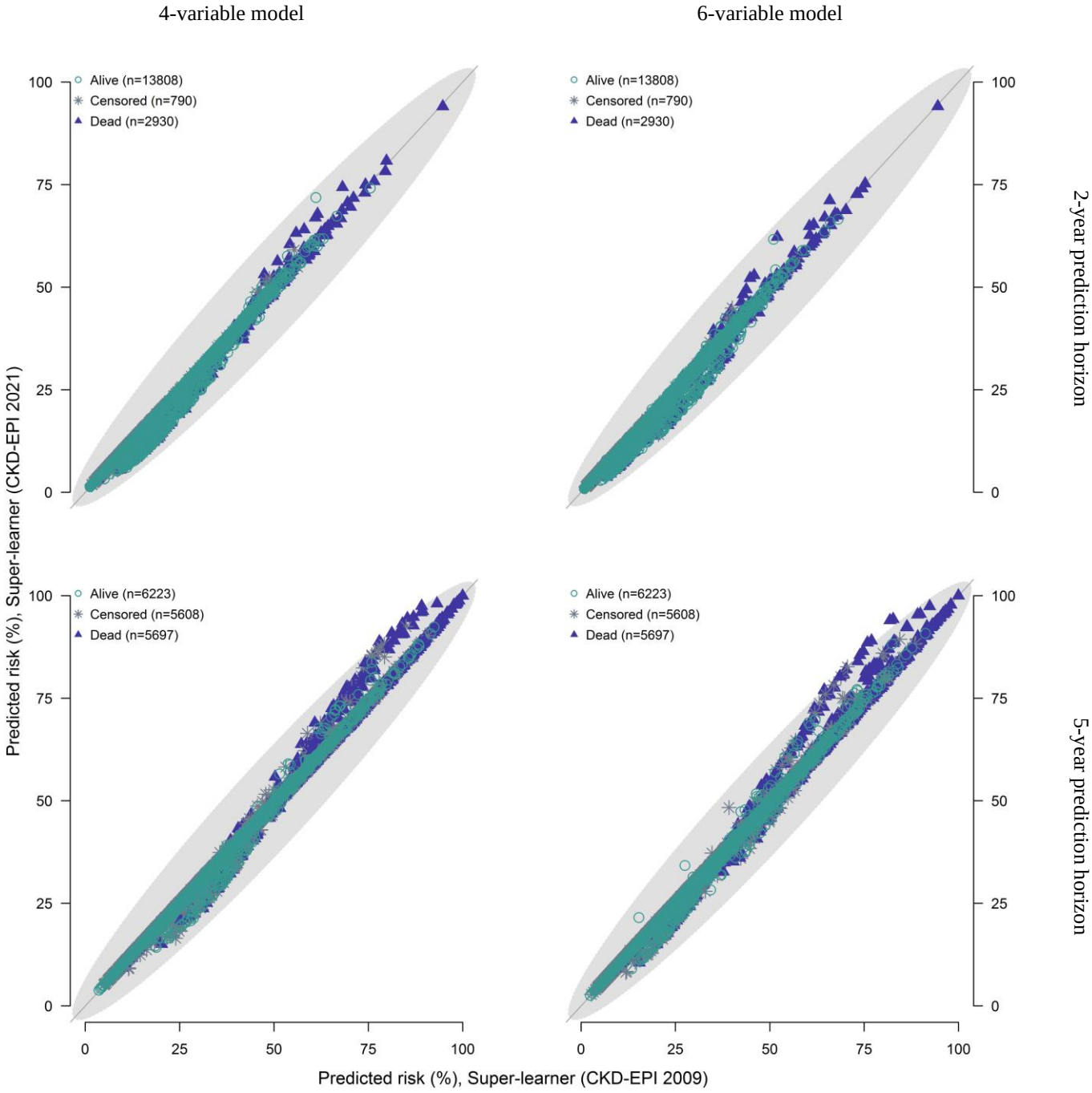
B. Scotland cohort



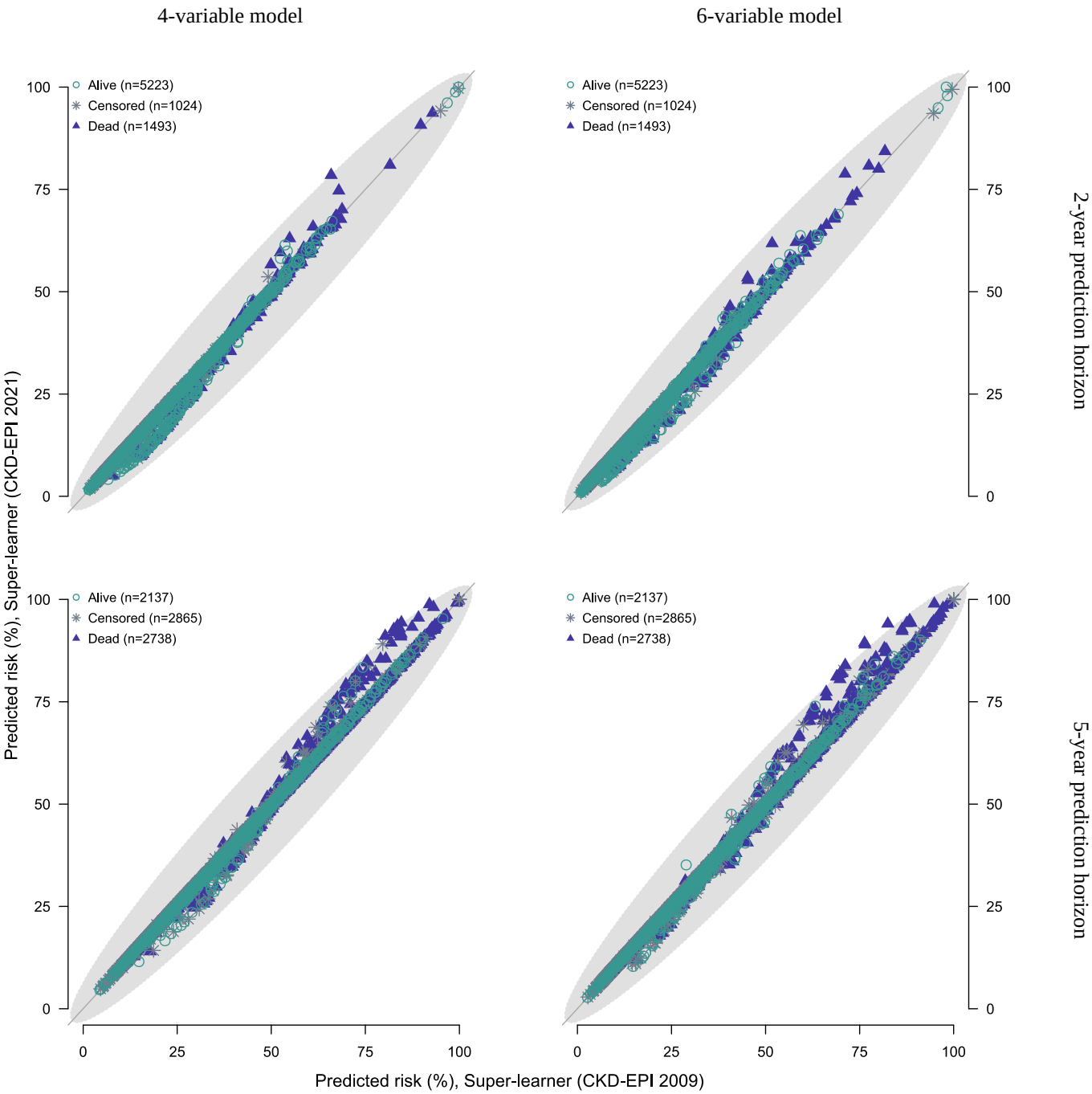
Legend: Scatterplot of predicted risks of kidney failure by GFR estimation formula (CKD-EPI 2009 vs 2021) from the 4-variable super-learner (left panels) and 6-variable super-learner (right panels), at prediction time horizon (2 years, top, for the G4 cohort and 5 years, bottom, for the G3bG4 whole cohort). Each point indicates the status of an individual at the prediction time horizon: alive without kidney failure, censored (unknown status), kidney failure, and competing risk (death without kidney failure). The gray-shaded region in each plot indicates the area of clinically meaningless risk difference (set at the pre-specified value of 10%). The KDpredict super-learner models were trained in Alberta, Canada, and applied in Denmark and Scotland.

Figure S19: Agreement between individual 2- and 5-year risk predictions of death (eGFR calculated with EPI 2009 vs EPI 2021 formula)

A. Denmark cohort

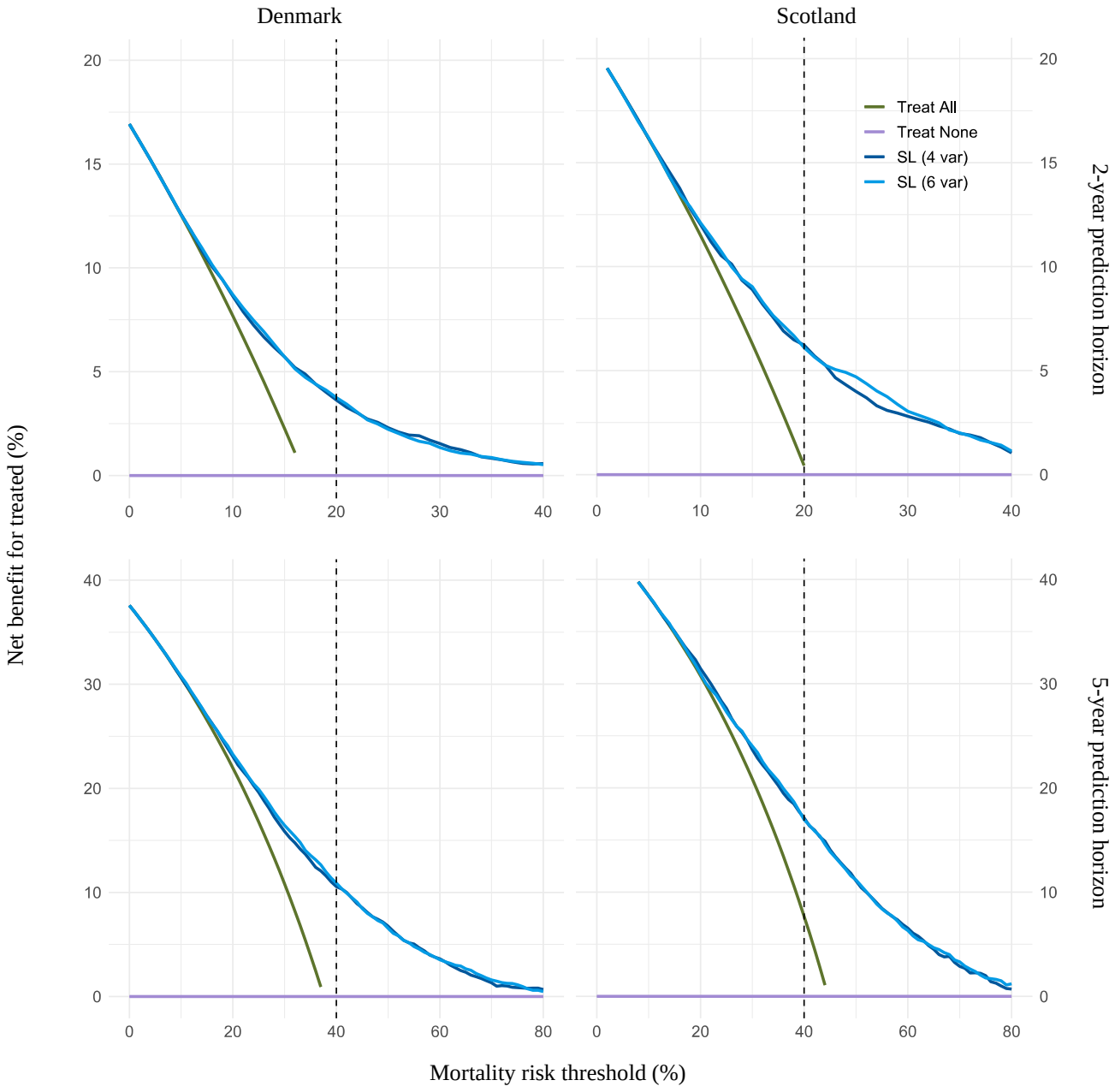


B. Scotland cohort



Legend: Scatterplot of predicted risks of death by GFR estimation formula (CKD-EPI 2009 vs 2021) from the 4-variable super-learner (left panels) and 6-variable super-learner (right panels), at prediction time horizon (2 years [top] and 5 years [bottom] for the G3bG4 whole cohort). Each point indicates the status of an individual at the prediction time horizon: alive, censored (unknown status), or dead. The gray-shaded region in each plot indicates the area of clinically meaningless risk difference (set at the pre-specified value of 10%). The KDpredict super-learner models were trained in Alberta, Canada, and applied in Denmark and Scotland.

Figure S20: Decision curve analysis of mortality



Legend: Net benefit of using different clinical strategies for 2- and 5-year risk prediction of mortality. SL (4 var) and SL (6 var), super-learner with 4 and 6 variables. Decision curves are presented by testing site (Denmark, left panels, and Scotland, right panels) and prediction time horizon (2 years top, and 5 years bottom). In decision curve analysis, "treated" is used in general sense to indicate the intervention decisions informed by different clinical strategies: treat all as if all would experience the event, treat none, treat based on alternative models. Decision curve analysis calculates the 'net benefit' by putting harm (false positive) on the same scale as benefit (true positive). To achieve this, false positive rates are multiplied by an exchange rate (how many false positives are worth one true positive) defined by a risk threshold. For mortality no risk thresholds exist for intervention decisions. Values of 20% and 40% at 2 and 5 years were prespecified for this analysis.

Supplementary References

1. Hemmelgarn BR, Clement F, Manns BJ, et al. Overview of the Alberta Kidney Disease Network. *BMC Nephrol* 2009;10:30. doi: 10.1186/1471-2369-10-30
2. Jensen SK, Heide-Jorgensen U, Vestergaard SV, et al. Routine Clinical Care Creatinine Data in Denmark - An Epidemiological Resource for Nationwide Population-Based Studies of Kidney Disease. *Clin Epidemiol* 2022;14:1415-26. doi: 10.2147/CLEP.S380840
3. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol* 2019;11:563-91. doi: 10.2147/CLEP.S179083
4. Sawhney S, Tan Z, Black C, et al. Validation of Risk Prediction Models to Inform Clinical Decisions After Acute Kidney Injury. *Am J Kidney Dis* 2021;78(1):28-37. doi: 10.1053/j.ajkd.2020.12.008 [published Online First: 2021/01/12]
5. 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(6):e2112828. doi: 10.1001/jamanetworkopen.2021.12828 [published Online First: 2021/06/09]
6. Liu P, Quinn RR, Lam NN, et al. Accounting for Age in the Definition of Chronic Kidney Disease. *JAMA Intern Med* 2021;181(10):1359-66. doi: 10.1001/jamainternmed.2021.4813 [published Online First: 2021/08/31]
7. Vestergaard SV, Christiansen CF, Thomsen RW, et al. Identification of Patients with CKD in Medical Databases: A Comparison of Different Algorithms. *Clin J Am Soc Nephrol* 2021;16(4):543-51. doi: 10.2215/CJN.15691020 [published Online First: 2021/03/13]
8. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009;150(9):604-12. doi: 10.7326/0003-4819-150-9-200905050-00006
9. 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(19):1737-49. doi: 10.1056/NEJMoa2102953 [published Online First: 2021/09/24]
10. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney inter., Suppl. 2013; 3: 1-150.
11. Tangri N, Stevens LA, Griffith J, et al. A predictive model for progression of chronic kidney disease to kidney failure. *JAMA* 2011;305(15):1553-59.
12. Tangri N, Grams ME, Levey AS, et al. Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis. *JAMA* 2016;315(2):164-74. doi: 10.1001/jama.2015.18202 [published Online First: 2016/01/13]
13. Pasternak M, Liu P, Quinn R, et al. Association of Albuminuria and Regression of Chronic Kidney Disease in Adults With Newly Diagnosed Moderate to Severe Chronic Kidney Disease. *JAMA Netw Open* 2022;5(8):e2225821. doi: 10.1001/jamanetworkopen.2022.25821
14. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013;22(3):278-95. doi: 10.1177/0962280210395740 [published Online First: 2011/01/12]
15. Tonelli M, Wiebe N, Fortin M, et al. Methods for identifying 30 chronic conditions: application to administrative data. *BMC Med Inform Decis Mak* 2015;15:31. doi: 10.1186/s12911-015-0155-5 [published Online First: 2015/04/18]
16. Al-Wahsh H, Lam NN, Quinn RR, et al. Calculated versus measured albumin-creatinine ratio to predict kidney failure and death in people with chronic kidney disease. *Kidney Int* 2022 doi: 10.1016/j.kint.2022.02.034 [published Online First: 2022/04/11]
17. 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. *Annals of internal medicine* 2020;173(6):426-35.
18. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441. doi: 10.1136/bmj.m441 [published Online First: 2020/03/20]

19. KDIGO. Kidney Disease Improving Global Outcomes. Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney Int* 2012(3):1–150.
20. Laan MJvd, Polley EC, Hubbard AE. Super Learner. *Statistical Applications in Genetics and Molecular Biology* 2007;6(1) doi: doi:10.2202/1544-6115.1309
21. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* 2001;16(3):199-231, 33.
22. Ozenne B, Lyngholm Sørensen A, Scheike T, et al. riskRegression: Predicting the Risk of an Event using Cox Regression Models. *The R Journal* 2017;9(2):440-60. doi: 10.32614/RJ-2017-062
23. Ishwaran H, Gerds TA, Kogalur UB, et al. Random survival forests for competing risks. *Biostatistics* 2014;15(4):757-73. doi: 10.1093/biostatistics/kxu010
24. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *The Annals of Applied Statistics* 2008;2(3):841-60. doi: 10.1214/08-AOAS169
25. Gerds TA, Kattan M.W. Medical risk prediction models: With ties to Machine Learning: Chapman and Hall/CRC, 2021.
26. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Annals of internal medicine* 2009;150(9):604-12.
27. Efron B, Tibshirani R. Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association* 1997;92(438):548-60. doi: 10.2307/2965703
28. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical chemistry* 2008;54(1):17-23.
29. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine* 2014;33(18):3191-203. doi: 10.1002/sim.6152
30. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research* 2018;2(1):7. doi: 10.1186/s41512-018-0029-2
31. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med* 2014;33(18):3191-203. doi: 10.1002/sim.6152
32. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;18(17-18):2529-45. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5
33. 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 Int* 2018;93(6):1442-51. doi: 10.1016/j.kint.2018.01.009 [published Online First: 2018/04/02]
34. 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. doi: 10.1136/bmj.i6 [published Online First: 2016/01/27]
35. 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. doi: 10.1186/s12911-015-0155-5