Predicting outcomes in patients with kidney disease: methodology and clinical applications

Ramspek, C.L.

5

Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models

Chava L. Ramspek, Lucy Teece, Kym I.E. Snell, Marie Evans, Richard D. Riley, Maarten van Smeden, Nan van Geloven, Merel van Diepen

Int J Epidemiol. 2021; Dec 17;dyab256 (online ahead of print)

Abstract

Background. External validation of prognostic models is necessary to assess the accuracy and generalizability of the model to new patients. If models are validated in a setting where competing events occur, these competing risks should be accounted for when comparing predicted risks to observed outcomes.

Methods. We discuss existing measures of calibration and discrimination which incorporate competing events for time-to-event models. These methods are illustrated using a clinical data-example concerning the prediction of kidney failure in a population with advanced chronic kidney disease (CKD), using the guideline-recommended Kidney Failure Risk Equation (KFRE). The KFRE was developed using Cox regression in a diverse population of CKD patients and has been proposed for use in patients with advanced CKD in whom death is a frequent competing event.

Results. When validating the 5-year KFRE with methods that account for competing events, it becomes apparent that the 5-year KFRE considerably overestimates the real-world risk of kidney failure. The absolute overestimation was 10 percentage points on average and 29 percentage points in older high-risk patients.

Conclusion. It is crucial that competing events are accounted for during external validation to provide a more reliable assessment of a model's performance in clinical settings where competing risks occur.

Glossary

Prediction horizon	The specified time period over which predictions are made, in our clinical validation this is 2 and 5 years.
Event of interest	The primary event that is being predicted, in our clinical validation study this is kidney failure.
Competing event	Any events that may preclude the primary event from happening, in this case death without kidney failure.
Absolute risk	The cumulative risk of the event of interest within the prediction horizon, given that patients may be censored and patients with a competing event will not experience the event of interest. This risk is also referred to as real-world risk, actual risk, crude risk or cumulative incidence. It can be calculated through a non-parametric cumulative incidence functions which is also termed Aalen-Johansen estimator.
Predicted risk	The risk predictions (output) from a prediction model over the specified prediction-horizon. In this study we assume the predicted risks are available, calculated from an existing model. The accuracy and precision of these predicted risks is evaluated in external validation.
Observed probability	The observed rate of the event of interest in the validation cohort, which is compared to the predicted risk. If there is no censoring and no competing events this is the proportion of patients who experience the primary event. If competing risks and censoring are present and the researcher wants to account for this, the observed probability for a group is the same as the absolute risk (detailed above).
"Accounting for competing events"	The use of methods that allow patients to fail from competing events. These patients are retained in the dataset but dealt with using assumptions in a way that precludes them from experiencing the event of interest after the competing event, thereby differing from the assumptions for patients censored due to loss to follow-up or other reasons.
"Ignoring competing events"	Using statistical methods with inappropriate assumptions concerning competing events, most often by assuming no competing risks or that competing risks could be eliminated.

1. Introduction

Prognostic models have rapidly become an integral part of medical practice. As clinical care moves towards individualized monitoring, decision-making, and treatment, it is imperative to collect information on an individual's risk profile and many prognostic models have been developed.(1-3) External validation of prognostic models is a crucial step to assess the accuracy and generalizability of the model but may present various methodological challenges including the occurrence of competing events.(4)

Competing events prohibit patients from experiencing the prognostic outcome of interest, and often occur when studying high-risk interventions, long prediction-horizons, or cause-specific mortality. For instance, the prediction of kidney failure in patients with advanced chronic kidney disease (CKD) is complicated due to patients dying from other causes before they can develop kidney failure. A conventional time-to-event regression model (such as a standard Cox model) that predicts an individual's risk of kidney failure would censor all patients with incomplete follow-up in the same manner, including patients with competing events (death). Such a model would therefore overestimate the absolute risk of kidney failure.(5-7) The overestimation of risks due to unaccounted competing events can result in counterintuitive and misleading prognostication. For instance, in a population with kidney failure, the 5-year risk of cardiovascular death and the 5-year risk of non-cardiovascular death sum to 107% when calculated separately without correctly accounting for competing events.(8) The calculated probabilities are hypothetical risks assuming no patient dies from the other cause. Though there are exceptions, 'the risk assuming no occurrence of death' ordinarily has little clinical relevance. In this study, we thus assume that researchers aim to estimate the absolute risk of prognostic outcomes in a real world setting in which competing events occur.

The importance of using appropriate competing risk modelling techniques (such as Fine & Gray subdistribution regression models or combined cause-specific Cox models) for prognostic model development is increasingly recognized.(6, 9-16) Nevertheless, most clinical time-to-event prognostic tools are developed using conventional regression models.(9, 17, 18) Therefore, it is important to recognize that the influence of competing events can also be evaluated during external validation, as will be illustrated in this article. By doing so, existing time-to-event models can be validated in settings in which competing events may be more or less frequent than the development setting. This paper was inspired by a recent publication from our research group in which existing kidney failure models were validated while accounting for the competing risk of death.(19) In this process many lessons on involved statistics and interpretation of results were learnt which we hope to share.

The aim of this paper is to draw attention to the importance of externally validating time-to-event prognostic models in a manner that appropriately accounts for competing events. First, we concisely discuss the technicalities of assessing performance measures in a competing risk setting. Secondly, we provide a real-data example in which we externally validate an existing prognostic model of kidney failure in patients with advanced CKD. This

example illustrates the effects of competing events on measures of prognostic performance and details how such analyses can shift clinical conclusions considerably.

2. Predictive performance at external validation

In this paper we assume that the time-to-event model of interest has already been developed and may or may not have accounted for competing risks. Secondly, we assume that the aim is to validate this model in a setting in which competing events occur, and that clinicians and patients want individualised absolute risk predictions that reflect this. Finally, we assume a specified prediction-horizon for which validation is of interest.

External validation of a prognostic model assesses the accuracy of predictions made by the model in individuals that were not used to develop the model.(20) Important elements of prognostic model performance are assessed by comparing how well the predicted risks agree with the observed outcomes (calibration) and how well predictions separate patients who will and will not experience the outcome of interest (discrimination). We now discuss existing measures of calibration and discrimination which incorporate competing events for time-to-event models. The supplementary material includes a more in-depth explanation on these various methods and we have provided a GitHub repository (in collaboration with authors from a STRATOS initiative guidance paper) with available R-code on how to validate a competing risk model.

Calibration

Calibration of predicted and observed outcomes can be assessed through calibration-in-the-large (overall calibration) and visualised using calibration plots.(21) When dealing with competing events it is key that the observed probability is calculated in a way that accurately accounts for the competing events and thereby represents the absolute risk of the event of interest.

Calibration-in-the-large can be assessed by comparing the average predicted risk for the outcome to the observed probability at the prediction horizon. Dividing the observed probability by the average predicted probability gives the O/E ratio. The average predicted risk is known, since we assume all individual predicted risks according to the existing prediction model are given. In the case of censoring, the non-parametric Kaplan-Meier estimator is often used to calculate the observed probability. However, in the presence of competing events, the KM estimate will overestimate the absolute risk of the event of interest.(8, 22) A more appropriate method to calculate the observed outcome probability in the presence of competing events is the non-parametric cumulative incidence function (CIF).(23) Calculating the CIF is similar to using the KM method, but quantifies the risk for the event of interest and competing events; all of which increase over time. Using the CIF, patients who experience a competing event are no longer at risk of experiencing the outcome of interest, and the probability of the outcome of interest is scaled by the

cumulative probability of experiencing any event. No assumptions are needed on the independence of competing events and the outcome.(6, 8)

In calibration plots, the predicted and observed outcome probabilities are plotted against each other to visualize their agreement. Often the cohort is divided into subgroups based on quantiles of predicted risks. The average predicted and observed outcome probabilities for each subgroup can be computed (accounting for competing events as described above) and plotted. This approach has been criticized as the categorization is arbitrary and can lead to loss of precision and misleading results.(24) It is therefore recommended to include a smoothed curve in the calibration plot. In the presence of censoring this smoothed curve is often based on pseudo-values. In the presence of competing events, this smoothed curve can be obtained using pseudo-values, as described by Gerds et al.(25) By using these pseudo-values which are based on cumulative incidence estimates, the model calibration is estimated over the full range of predicted probabilities.

Discrimination

Discrimination examines the model's ability to distinguish between those who will experience the outcome of interest from those who will not and is based on the ranked order of predicted risks.(26) For survival data, Harrell's C-index is the most frequently reported measure of discrimination, which is the proportion of all examinable pairs in which the individual with the highest predicted risk is observed to experience the outcome sooner than the other individual.(24) A C-index of 1 is perfect discrimination and 0.5 is equivalent to chance. Censored patients are treated as if they might still experience the outcome in the future which is an incorrect assumption in the case of censoring due to a competing event.(27)

In the presence of competing events, various methods to calculate a C-index are available, some of which are referenced.(11, 13, 28) In the case of complete outcome data (no or very few patients are lost to follow-up), a simple adaptation of Harrell's C-index as proposed by Wolbers et al. can be employed.(11) Instead of censoring patients who experience a competing event, these patients are retained in the risk set whilst setting their follow-up time to infinity (or the prediction horizon), thus indicating that they will never experience the event of interest. In the case of only administrative censoring, also termed 'censoring complete', an adaptation of the Wolbers' approach can be used in which patients with the competing event are censored at the administrative censoring date (instead of infinity).(29, 30) In the case of informative censoring more suitable methods are available; some of which have been adapted for competing risks settings, using inverse probability of censoring weighting (IPCW).(13, 28, 31, 32) In IPCW a pseudo-population that would have been observed if no censoring occurred, is created. This pseudo-population contains only patients who are followed until they experience either the event of interest, a competing event or the end of follow-up. This is done by upweighting patients who are similar to censored patients but remain in the study (under the assumptions of exchangeability, consistency and positivity). Royston and Sauerbrei's D statistic is a measure of prognostic

separation.(33) It can be interpreted as the coefficient (log hazard ratio) for comparing two equal-sized prognostic groups, created by dichotomizing the model's linear predictor estimates in the cohort at the median value.(34) Higher values of the D statistic represent greater separation between the survival curves for these prognostic groups. To calculate the D statistic in an external validation study, the linear predictors (for each individual) from the prognostic model are ranked and scaled. The scaled ordering of the linear predictors is then entered in a new regression model with the event of interest as the outcome; the resulting regression coefficient is the D statistic. In an external validation of a time-to-event model, the scaled linear predictor values are generally entered into a new Cox model. To adapt this measure to a setting with competing events in an external validation, the Cox model can be replaced by a Fine & Gray regression model.(35) The D statistic can be transformed to the proportion of explained variation: R^2_D.(36) This measure indicates how much of the observed variation in the outcome is explained by the prognostic model.

3. Real-data illustration: predicting kidney failure in advanced chronic kidney disease patients from the Swedish Renal Registry

Rationale

Predicting kidney failure in advanced CKD patients is of interest for timely preparation of dialysis and transplantation, adequate monitoring of patients, possible referral back to primary care and informing patients of their likely prognosis. As the rate of progression to kidney failure highly varies between individuals, prognostic models have been proposed for use in clinical practice. The Kidney Failure Risk Equation (KFRE) is a prognostic model that was developed to predict kidney failure in patients with CKD stage 3-5 who were referred to a nephrologist.(32) It was later externally validated and updated in a large meta-analysis and is recommended for use in international medical guidelines.(37-39)

Cox proportional hazards models were used in KFRE model development and external validation studies, meaning patients who died before experiencing kidney failure were censored.(32, 39-43) This means the predicted outcome is the risk of kidney failure in a setting in which patients are prevented from dying at least until kidney failure occurs. This risk is, however, not defined as such in the study. Instead, the predicted risk is presented as the absolute risk of kidney failure which is more clinically relevant and conducive towards medical decision making. In the KFRE development study, the use of a competing risk model was explored as sensitivity analysis, but not published as the predicted risks were deemed similar to those from the Cox model.

In this clinical illustration the aim is to externally validate the KFRE in two ways, first using methods that are fitting for a Cox prediction model and treat patients with a competing event the same as any other censored patient (similar to the development study).

Secondly, we will use methods described previously to account for competing events in order to validate how well the KFRE predicts the real-world risk of kidney failure.

Methods

The KFRE includes the four following predictors: age, sex, eGFR and urine albumin-to-creatinine ratio. The outcome kidney failure is defined by the initiation of dialysis or kidney transplantation within 2 or 5 years. The full prediction formulas are provided in the development studies and are also shown in our supplement.

Patients were included from the Swedish Renal Registry (SRR), an ongoing registry of chronic kidney disease patients capturing 98% of the nephrology clinics in Sweden.(44, 45) Patients who entered the registry between January 1st 2012 and June 30th 2018 were included. The analysis was restricted to patients aged 18 years or older with an estimated glomerular filtration rate (eGFR) between 8 and 30 ml/min/1.73m^2. The eGFR is a measure of kidney function; below 30 indicates advanced CKD. Time zero (moment of prediction) was inclusion in the SRR which is generally the first referral to a nephrologist.

Results

In total, 13,489 patients were included in our analysis of whom 1,818 (13%) developed kidney failure (the outcome of interest) within 2 years and 2,764 (20%) within 5 years. Slightly more patients died without experiencing kidney failure; 2,158 (16%) within 2 years and 3,357 (25%) within 5 years. No patients were lost to follow-up. All patients were administratively censored on June 30th 2018. The median follow-up was 1.7 years and the maximum 6.7 years. In total, 3,548 patients (26%) were administratively censored within 2 years and 6,410 patients (48%) within 5 years. For each individual, the predicted 2 and 5-year risks were calculated using the KFRE formulae. Missing predictors were imputed using the R-package mice.(46) For the illustrative purposes of this article we used a single imputed dataset for all analyses, more information on the imputation and baseline data is shown in the supplemental material.

The difference between observed outcome probabilities of kidney failure, death, and event-free survival, calculated using the KM and CIF methods, are shown in a stacked histogram (Figure 1) and in cumulative incidence curves (Figure 2). At two years, the KM risk for death and kidney failure are both 2 percentage points higher than when calculated with the CIF, resulting in a total risk of 104%. At 5 years, the sum of the risks using KM increases to 120%. Risks based on the CIF method always sum to 100%.

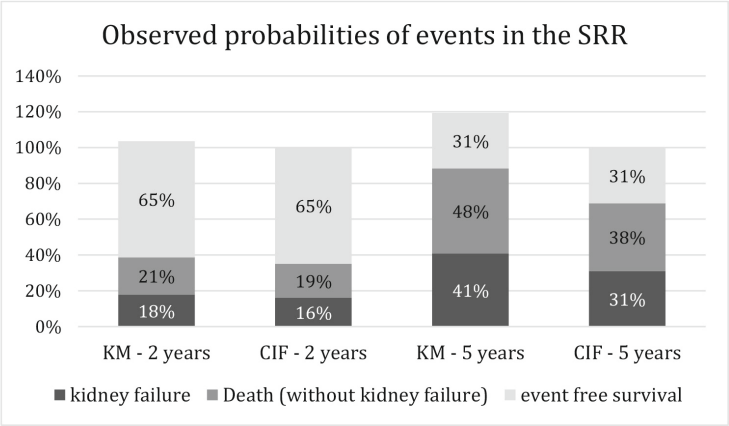

Figure 1: Differences between KM and CIF estimates of the observed outcome probabilities in the presence of competing events.

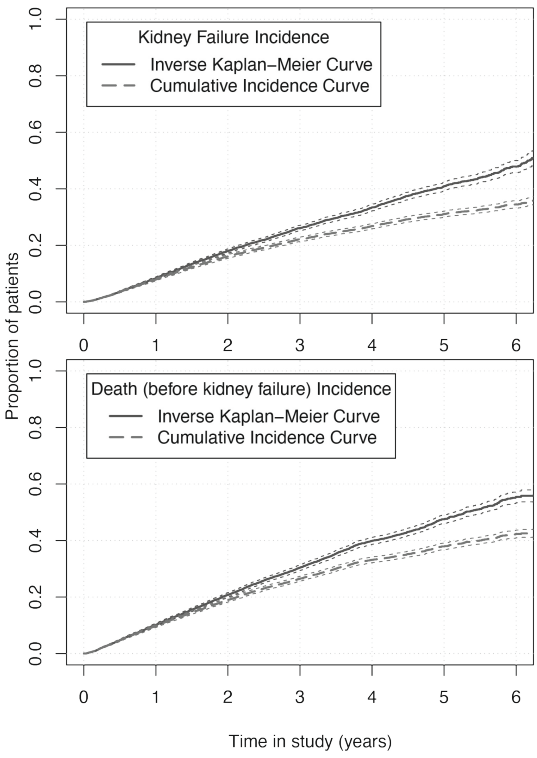

Figure 2: One minus Kaplan-Meier curves and cumulative incidence curves of the observed outcome probabilities in the SRR for kidney failure and Death. For illustrative purposes patients who experienced kidney failure were censored or regarded as competing event in the lower plot.

To assess the calibration-in-the-large, the observed kidney failure outcome probabilities based on KM and CIF were compared to the model's average predicted risk of kidney failure (Table 1). The 2-year KM and 5-year KM outcome probabilities are both similar to the average predicted probability. When we consider the competing risk of death using the CIF, the observed 2-year probability of kidney failure is slightly lower but still similar to the model's average predicted risk with an O/E of 0.94 (95% CI: 0.91-0.98). The 5-year observed probability however is almost 10 percentage points lower than the predicted risk, with a corresponding O/E of 0.76 (95% CI: 0.74-0.78). Similar results are seen in the calibration plot using KM and CIF. In Figure 3a, the 2-year calibration curves for both methods are quite similar. In Figure 3b, the calibration plot for the 5-year KFRE is shown. When calculating observed probability with the standard KM method, calibration appears to be excellent. However, when we take the competing risk of death into account, the KFRE appears to considerably overpredict the actual proportion of patients with kidney failure, particularly in high-risk patients. Out of the tenth of patients with an average predicted 5-year kidney failure risk of 81%, only 58% (95% CI: 56%-61%) experienced kidney failure.

For model discrimination, the differences are less pronounced between accounting for competing risks or not (Table 1). When patients who die are censored, the standard Harrell's C-index is 0.829 for the 5-year KFRE. When these patients are no longer censored but set to the follow-up time they would have had if administratively censored (to indicate that patients who die will not experience kidney failure), the C-index is slightly lower: 0.814. The D statistic and explained variance also reflect that when competing risks are accounted for, the 5-year discrimination is slightly lower (Table 1).

Table 1: Calibration and discrimination results for external validation of the 2 and 5-year KFRE, in the entire validation cohort (n=13489). The external validation was performed in two manners, first by ignoring the competing risk of death by censoring these patients and using KM-estimates. Secondly, we validated the models whilst taking account of competing risks in all performance measures.

	KFRE 2-year model		**KFRE 5-year model**	
	Ignoring competing events by censoring	*Taking competing events into account*	*Ignoring competing events by censoring*	*Taking competing events into account*
Average predicted risk	17%	17%	41%	41%
Average observed probability (95% CI)	18% (17%-19%)	16% (15%-17%)	41% (40%-42%)	31% (30%-32%)
O/E ratio	1.06 (1.02-1.10)	0.94 (0.91-0.98)	1.00 (0.98-1.02)	0.76 (0.74-0.78)
C-index (95% CI)	0.840 (0.831-0.849)	0.834 (0.825-0.843)	0.829 (0.821-0.837)	0.814 (0.806-0.822)
D statistic (95% CI)	2.34 (2.25-2.42)	2.32 (2.20-2.43)	2.13 (2.06-2.19)	2.04 (1.95-2.14)
$\mathbf{R^2_D}$	57%	56%	52%	50%

Abbreviations: KFRE: kidney failure risk equation, O/E: observed/expected, CI: confidence interval

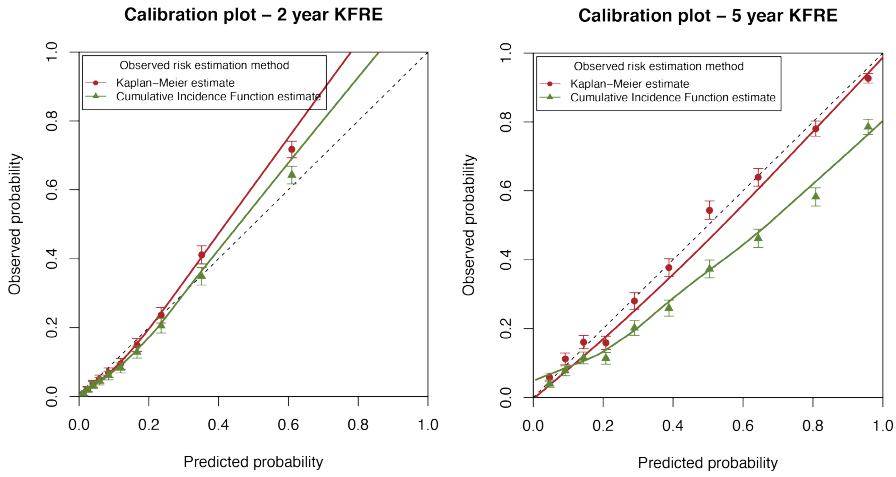

Figure 3a (left) and b (right): Calibration plots for external validation of the 2 and 5-year KFRE. The external validation was performed by using KM-estimates (ignoring competing risks) and by using a competing risks approach. The competing risks approach (green points and line) represents the model performance for the absolute kidney failure risk in a setting in which patients may die.

As death without kidney failure is more frequent in older CKD patients, we also validated the KFRE in a subgroup of patients who were 70 years or older (n=8654). These patients had a higher risk of death; 1064 patients (12%) experienced kidney failure within 5 years, whilst 2847 patients (33%) died without kidney failure. The median follow-up time was 1.7 years and the maximum follow-up time 6.5 years. All analyses were repeated in this subgroup and these results are shown in Table 2 and Figure 4. Overall, the differences between ignoring competing events and accounting for them are even more pronounced in this high-risk subgroup. These differences are larger for the 5-year model, and more apparent in measures of calibration than discrimination. For the 5-year model the O/E is 0.84 (95% CI: 0.81-0.87) when ignoring competing events and 0.57 (95% CI: 0.54-0.59) when accounting for them. The 10% of patients with the highest predicted 5-year risk (most right data-point in Figure 4b) have an average predicted risk of 89%. Without considering the competing risk of death, 81% (95% CI: 78%-83%) of them are expected to experience kidney failure. However, when accounting for competing events, we observe that only 52% (95% CI: 48%-55%) of these patients actually experience kidney failure.

Table 2: Calibration and discrimination results for external validation of the 2 and 5-year KFRE, in a subset of patients aged 70+ years (n=8654). The external validation was performed in two manners, first by ignoring the competing risk of death by censoring these patients and using KM-estimates. Secondly, we validated the models whilst taking account of competing risks in all performance measures.

	KFRE 2-year model		KFRE 5-year model	
	Ignoring competing events by censoring	*Taking competing events into account*	*Ignoring competing events by censoring*	*Taking competing events into account*
Average predicted risk	13%	13%	34%	34%
Average observed probability (95% CI)	11% (11%-12%)	10% (9%-10%)	28% (27%-29%)	19% (18%-20%)
O/E ratio (95% CI)	0.91 (0.86-0.96)	0.78 (0.73-0.83)	0.84 (0.81-0.87)	0.57 (0.54-0.59)
C-index (95% CI)	0.826 (0.810-0.841)	0.813 (0.797-0.828)	0.817 (0.803-0.830)	0.791 (0.778-0.805)
D statistic (95% CI)	2.23 (2.10-2.36)	2.04 (1.90-2.17)	2.09 (1.98-2.20)	1.75 (1.63-1.86)
R^2_D	54.3%	49.8%	51.1%	42.1%

Abbreviations: KFRE: kidney failure risk equation, O/E: observed/expected, CI: confidence interval

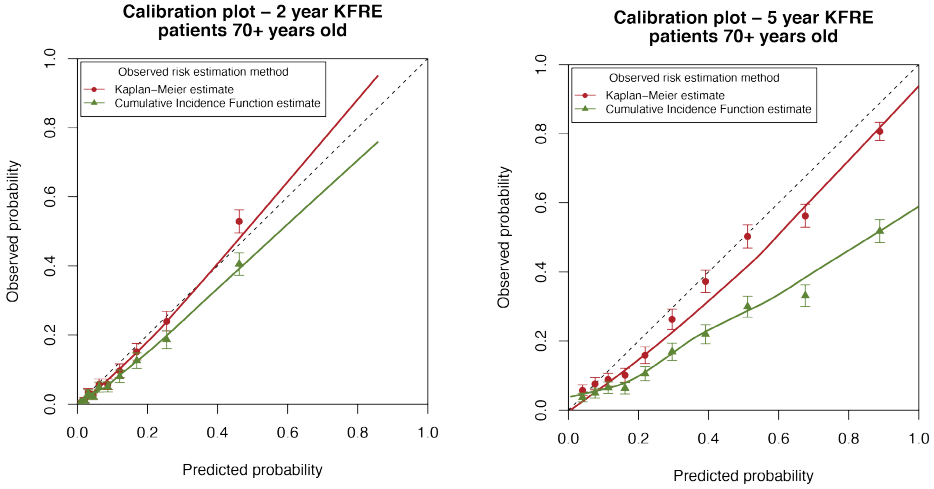

Figure 4a (left) and b (right): Calibration plots for external validation of the 2 and 5-year KFRE in a subset of older patients. The external validation was performed by using KM-estimates (ignoring competing risks) and by using a competing risks approach. The competing risks approach (green points and line) represents the model performance for the absolute kidney failure risk in a setting in which patients may die.

Conclusions

From the external validation of the KFRE in which we have taken the competing risk of death into account, we conclude that the 2-year KFRE adequately predicts the absolute risk of kidney failure in patients with advanced CKD. However, if we wish to interpret the kidney failure risk as kidney failure in a real-world setting with competing events, the 5-year KFRE is poorly calibrated and considerably overestimates the absolute risk of kidney failure. This overprediction is more pronounced in older patients. The difference between performance of the 2 and 5-year model can be attributed to a lower number of patients dying without kidney failure within 2-years. If clinicians interpret the 5-year KFRE estimate as the absolute kidney failure risk (instead of the hypothetical risk given no patient can die before kidney failure), the overestimation could lead to patients being unnecessarily prepared for dialysis (which includes vascular access surgery and frequent hospital visits). As the four variable 5-year KFRE substantially overpredicted kidney failure risk when considering the competing risk of death, this model is not recommended for use in patients with advanced CKD. An alternative model which accounts for competing events, such as the 4-year Grams model is recommended instead.(47) This model has recently been compared head-to-head with the KFRE in an external validation study and demonstrated superior performance when accounting for the competing risk of death.(19)

4. Discussion

In this paper, we highlighted the importance and implications of appropriately managing competing events during external validation. We provided explanation and tools on existing measures of calibration (O/E ratio and calibration plots) and discrimination (C-index, D statistic and R^2_D) that have been adapted to a competing risk setting.

The importance of competing event analyses has received increased attention in prognostic research.(6, 9-11, 17, 18) However, existing studies have mainly focussed on the importance of using competing risks methods in the development of prognostic models. It may well be that a prognostic model is developed in a setting with no or very few competing events, and therefore a valid representation of the absolute risk for that population. However, if that model is then validated in a different population in which competing events are more frequent, it is crucial that these competing events are appropriately managed in the external validation process.

The presence of competing events may influence all model performance measures, though in general the effect on absolute measures (calibration) is larger than on relative measures (discrimination). Researchers should carefully consider and select the risk they wish to predict; if a model censors patients that experience a competing event, the predicted risk is the hypothetical risk in a setting in which the competing event does not exist.(48) If death is the competing event, approaches that assume no competing risks will give a more extreme overestimation of the absolute risk in older populations and for longer

prediction-horizons, as shown in our data-example. This overestimation will be overlooked if conventional validation methods are used.

The predicted risk of prognostic models is crucial in regard to medical decision-making. For instance, the KFRE is proposed for use in timely preparation for dialysis and kidney transplantation. Predicted risks that are too high, may negatively influence clinical treatment decisions. External validation without accounting for competing risks, may lead to implementation of prognostic models that surreptitiously overpredict real-world outcomes and consequently result in overtreatment of patients.

The current study has a number of limitations. We have not developed any novel statistical approaches and do not provide information on how to adapt all available performance measures to a competing risk setting. Particularly measures of net benefit and decision-curve analysis were outside the scope of the current paper. Additionally, further research may focus on adapted measures of the calibration slope and integrated calibration index to a setting with competing events.(49) Furthermore, our data-example is based on a single dataset and some of the observed results may be attributable to sampling variability. In the future, a data-simulation study in which the outcome, competing event and censoring prevalence is varied, may provide more insight on how model performance is affected in different competing risks scenarios. Although the data example focussed on the validation of the KFRE which was developed using a Cox prognostic model, a strength of the current paper is that the discussed methods are applicable to other time-to-event models such as (flexible) parametric models, competing risks models or machine learning models such as random survival forests.

In conclusion, depending on the underlying clinical question, competing events may be crucial to consider when externally validating time-to-event prognostic models. If an existing prediction model has targeted the incorrect estimand, we can expect a poorer performance when validating this model while accounting for competing events (and thereby adjusting the estimand). Such external validation studies can help determine whether such models are transportable to a real-life setting in which competing events occur.

Acknowledgements

Daniele Giardiello and Edouard Bonneville are gratefully acknowledged for their work on the accompanying R-code.

References

1. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328.
2. de Jong Y, Ramspek CL, van der Endt VHW, Rookmaaker MB, Blankestijn PJ, Vernooij RWM, et al. A systematic review and external validation of stroke prediction models demonstrates poor performance in dialysis patients. J Clin Epidemiol. 2020;123:69-79.
3. Ramspek CL, de Jong Y, Dekker FW, van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. Nephrol Dial Transplant. 2020;35(9):1527-38.
4. Riley RD, van der Windt D, Croft P, Moons KG. Prognosis Research in Healthcare: concepts, methods, and impact: Oxford University Press; 2019.
5. Prentice RL, Kalbfleisch JD, Peterson AV, Jr., Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978;34(4):541-54.
6. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601-9.
7. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.
8. Verduijn M, Grootendorst DC, Dekker FW, Jager KJ, le Cessie S. The analysis of competing events like cause-specific mortality--beware of the Kaplan-Meier method. Nephrol Dial Transplant. 2011;26(1):56-61.
9. Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrology Dialysis Transplantation. 2013;28(11):2670-7.
10. Ravani P, Fiocco M, Liu P, Quinn RR, Hemmelgarn B, James M, 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 : JASN. 2019;30(11):2219-27.
11. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology (Cambridge, Mass). 2009;20(4):555-61.
12. Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. Biometrics. 2010;66(4):999-1011.
13. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics. 2014;15(3):526-39.
14. Zhang Z, Cortese G, Combescure C, Marshall R, Lee M, Lim HJ, et al. Overview of model validation for survival regression model with competing risks using melanoma study data. Ann Transl Med. 2018;6(16):325.
15. Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biom J. 2011;53(1):88-112.
16. Heyard R, Timsit J-F, Held L, consortium C-M. Validation of discrete time-to-event prediction models in the presence of competing risks. Biometrical Journal. 2020;62(3):643-57.
17. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Stat Med. 2012;31(11-12):1089-97.
18. Li L, Yang W, Astor BC, Greene T. Competing Risk Modeling: Time to Put it in Our Standard Analytical Toolbox. 2019;30(12):2284-6.
19. Ramspek CL, Evans M, Wanner C, Drechsler C, Chesnaye NC, Szymczak M, et al. Kidney Failure Prediction Models: A Comprehensive External Validation Study in Patients with Advanced CKD. Journal of the American Society of Nephrology. 2021;32(5):1174-86.
20. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
21. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016;74:167-76.

22. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. British Journal of Cancer. 2004;91(7):1229-35.
23. Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. The Annals of Statistics. 1988;16(3):1141-54.
24. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis: Springer; 2015.
25. 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.
26. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
27. Uno H, Cai T, Tian L, Wei LJ. Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models. Journal of the American Statistical Association. 2007;102(478):527-37.
28. Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32(30):5381-97.
29. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.
30. Geskus RB. Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring. Biometrics. 2011;67(1):39-49.
31. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105-17.
32. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. Jama. 2011;305(15):1553-9.
33. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23(5):723-48.
34. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
35. Teece L. Investigating the presence and impact of competing events on prognostic model research: Keele University; 2019.
36. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology. 2013;13(1):33.
37. Findlay A, Farrington K, Joosten H, Macias JF, Mooney A, Tattersall J, et al. Clinical Practice Guideline on management of older patients with chronic kidney disease stage 3b or higher (eGFR<45 mL/min/1.73 m2): a summary document from the European Renal Best Practice Group. Nephrology Dialysis Transplantation. 2017;32(1):9-16.
38. Chan CT, Blankestijn PJ, Dember LM, Gallieni M, Harris DCH, Lok CE, et al. Dialysis initiation, modality choice, access, and prescription: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2019;96(1):37-47.
39. Tangri N, Grams ME, Levey AS, Coresh J, Appel LJ, Astor BC, et al. Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis. Jama. 2016;315(2):164-74.
40. Lennartz CS, Pickering JW, Seiler-Mußler S, Bauer L, Untersteller K, Emrich IE, et al. External Validation of the Kidney Failure Risk Equation and Re-Calibration with Addition of Ultrasound Parameters. Clin J Am Soc Nephrol. 2016;11(4):609-15.
41. Grams ME, Li L, Greene TH, Tin A, Sang Y, Kao WH, et al. Estimating time to ESRD using kidney failure risk equations: results from the African American Study of Kidney Disease and Hypertension (AASK). Am J Kidney Dis. 2015;65(3):394-402.
42. Peeters MJ, van Zuilen AD, van den Brand JA, Bots ML, Blankestijn PJ, Wetzels JF. Validation of the kidney failure risk equation in European CKD patients. Nephrol Dial Transplant. 2013;28(7):1773-9.
43. Levin A, Rigatto C, Barrett B, Madore F, Muirhead N, Holmes D, et al. Biomarkers of inflammation, fibrosis, cardiac stretch and injury predict death but not renal replacement therapy at 1 year in a Canadian chronic kidney disease cohort. Nephrol Dial Transplant. 2014;29(5):1037-47.

44. Lundström UH, Gasparini A, Bellocco R, Qureshi AR, Carrero J-J, Evans M. Low renal replacement therapy incidence among slowly progressing elderly chronic kidney disease patients referred to nephrology care: an observational study. BMC Nephrology. 2017;18(1):59.
45. Methven S, Gasparini A, Carrero JJ, Caskey FJ, Evans M. Routinely measured iohexol glomerular filtration rate versus creatinine-based estimated glomerular filtration rate as predictors of mortality in patients with advanced chronic kidney disease: a Swedish Chronic Kidney Disease Registry cohort study. Nephrol Dial Transplant. 2017;32(suppl_2):ii170-ii9.
46. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010:1-68.
47. Grams ME, Sang Y, Ballew SH, Carrero JJ, Djurdjev O, Heerspink HJL, et al. Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate. Kidney International. 2018;93(6):1442-51.
48. van Geloven N, Swanson S, Ramspek CL, Luijken K, van Diepen M, Morris TP, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. European Journal of Epidemiology. 2020.
49. Austin PC, Harrell Jr FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med. 2020;39(21):2714-42.

Supplemental Material for Chapter 5

Statistical details and code

A GitHub repository is available at https://github.com/survival-lumc/ValidationCompRisks. This GitHub page accompanies a more in depth STRATOS statistical guideline on available methods to validate competing risk models from our co-author Nan van Geloven. This STRATOS statistical guideline is still in preparation at the current time of article submission. The Prediction_CSC.md in depth markdown document with script provides R-code for the validation measures discussed in our main manuscript as detailed below.

1. Calibration-in-the-large

The O/E ratio using non-parametric cumulative incidence functions to calculate the observed probability is a measure of calibration-in-the-large or overall calibration. R-code to calculate this is shown in section 2.1.3 of the in-depth GitHub repository (in-depth markdown document). The corresponding confidence interval can be calculated according to a method by Debray which is also included in the R-code.[1]

2. Calibration plot

For the quantiles in a calibration plot the methods detailed under calibration-in-the-large can be used in subgroups. For a smoothed curve a non-parametric estimation method has been proposed using pseudo-observations. The R-code is provided in section 2.1.1.1 of the GitHub repository. The pseudo-observation for a particular patient is calculated by taking the weighted difference between the cumulative incidence estimate at the prediction horizon based on all patients and the same value leaving that patient out. This pseudo-observation is between 0 and 1 and functions as the observed probability for an individual patient. The advantage is that censored patients (who don't have an event indicator) do have a pseudo-observation. After transforming the data into pseudo-observations, a smooth curve of actual risks can be obtained using a nearest-neighbor smoother. This smoother averages the pseudo-observations within a small interval using a rolling bandwidth along the observed distribution of the risk estimates.[2]

3. C-statistic

In the case of complete outcome data, an adaptation of Harrell's C-index as proposed by Wolbers et al. can be employed.[3] Instead of censoring patients who experience a competing event, these patients are retained in the risk set whilst setting their follow-up time to infinity (or the prediction horizon), thus indicating that they will never experience the event of interest. Pairs where one individual has the primary event (within the prediction horizon) and the other has the primary event later or experiences

a competing event can be compared. The R-code is provided in section 2.2.1 of the GitHub repository. The C-index is influenced by the censoring distribution and this is particularly problematic when this censoring distribution depends heavily on other covariates.[4] When pairing cases with non-cases, Harrel's C-index cannot evaluate a pair in which the non-case is censored at an earlier time-point than the case. These non-evaluable pairs are ignored and this may induce bias.[4] More appropriate methods for calculating the C-index in time-to-event data with independent censoring have been developed. Most of these methods use inverse probability censoring weights (IPCW). In IPCW a pseudo-population that would have been observed if each patient were a complete-case, is created. A complete-case is an individual that has either experienced the event of interest, a competing event or is still at risk at the prediction-horizon. Complete-case patients are weighed inversely to their probability of having their particular outcome. In other words, patients who were not likely to remain in follow-up (but did), are up-weighted. To minimize bias in an external validation study of a time-to-event model with a considerable number of patients with dependent censoring, we advise to use IPCW estimates of the C-index.[5]

4. Royston-Sauerbrei D statistic and R^2_D

R-code for Royston and Sauerbrei's D-statistic as measure of prognostic separation and the R^2_D can be found in section 2.2.3. To calculate this, each individual's linear predictor value is ordered and the corresponding rankits (standard normal order statistics) are calculated and scaled by a factor $k = \sqrt{8/\pi}$. The scaled rankits are regressed on the outcome using a Fine & Gray model in the case of competing events. The resulting regression coefficient is the D-statistic. The D-statistic can be scaled to the log relative hazard scale to calculate the R^2_D. The D-statistic and R^2_D rely on a proportional hazards assumption and the assumption that the underlying linear predictor values are normally distributed (normality assumption). [6 7]

KFRE model

The KFRE Web calculator can be found at: https://kidneyfailurerisk.com/. To compute predicted risks, eGFR was calculated with the CKD-Epi formula. ACR is in mg/g, serum albumin in g/dL, phosphate in mg/dL, bicarbonate in mEq/L, calcium in mg/dL. The following non-North America formulas were used (as provided in the KFRE eAppendix 2 of the meta-analysis and update paper).[43] KFRE 4 variable 2-year probability = 1 – 0.9832 ^ exp (-0.2201 × (age/10 – 7.036) + 0.2467 × (male – 0.5642) – 0.5567 × (eGFR/5 – 7.222) + 0.4510 × (logACR – 5.137)). KFRE 4 variable 5-year probability = 1 – 0.9365 ^ exp (-0.2201 × (age/10 – 7.036) + 0.2467 × (male – 0.5642) – 0.5567 × (eGFR/5 – 7.222) + 0.4510 × (logACR – 5.137))

Multiple imputation and baseline data

For the purpose of this illustration a single multiple imputation was used with 5 iterations, instead of multiple imputations. However, all suggested methods can be applied on multiply imputed data, though for calibration choices will have to be made on whether to use the predicted risk from one of the imputed datasets at random or combine predicted risks from all imputed sets for an overall mean predicted risk per individual. ACR was the only predictor with missing values (42%). Our single imputation included the following variables as predictors at time zero to impute ACR: diabetes, hypertension, cardiovascular disease, blood pressure, albumin, calcium, phosphate, potassium, bicarbonate, eGFR, age, gender, log(ACR) at 6 months, log(ACR) at 12 months, kidney failure & time to kidney failure, death & time to death.

Baseline table of the SRR population. Continuous baseline characteristics are presented as mean values with standard deviations or median values with interquartile ranges when not normally distributed.

	Missing	Total n = 13489	No kidney failure within 5 years n=10725	Kidney failure within 5 years n=2764
Age (year)	0%	74.3 (65.7-81.2)	76.0 (68.5-82.2)	66.6 (53.8-74.2)
Sex (% male)	0%	61.3%	60.0%	66.4%
Primary Kidney Disease (%)	0%			
Diabetes mellitus		21.5%	18.8%	32.1%
Glomerular disease		6.9%	5.4%	12.7%
Hypertension		30.2%	32.7%	20.4%
Other		41.4%	43.1%	34.8%
Comorbidities (%)				
Congestive heart failure	0%	21.0%	23.3%	12.1%
Cardiovascular disease (other)	0%	21.3%	23.3%	13.5%
Hypertension	0%	73.2%	75.0%	66.1%
Diabetes mellitus	0%	36.4%	35.0%	41.9%
Laboratory parameters				
eGFR (MDRD) (ml/min/1.73m^2)	0%	21.9 (5.7)	22.9 (5.3)	18.2 (5.6)
ACR urine (mg/mmol)	41.8%	36 (7 - 155)	24 (5 - 101)	175 (57 - 340)
Serum Albumin (g/L)	6.9%	36 (5.2)	37 (34 - 40)	35 (32 - 39)
Serum Creatinine (µmol/L)	0%	227 (194-278)	232 (65)	306 (97)
Serum Calcium (mmol/L)	13.4%	2.29 (0.29)	2.31 (0.16)	2.24 (0.18)
Serum Phosphate (mmol/L)	9.0%	1.30 (0.29)	1.26 (0.27)	1.44 (0.32)
Serum Bicarbonate (mmol/L)	73.4%	22 (3.4)	23 (3.4)	22 (3.2)

Baseline table of the SRR population. Continued.

	Missing	Total n = 13489	No kidney failure within 5 years n=10725	Kidney failure within 5 years n=2764
Serum Potassium (mmol/L)	54.4%	4.43 (0.55)	4.41 (0.53)	4.50 (0.59)
Clinical parameters				
Body-mass index (kg/m^2)	33.5%	28.3 (6.0)	28.2 (6.0)	28.5 (6.3)
Systolic Blood pressure (mmHg)	5.7%	141 (22)	139 (22)	147 (22)
Diastolic Blood Pressure (mmHg)	5.8%	77 (12)	76 (12)	81 (13)

References

1. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *Bmj* 2017;356:i6460. doi: 10.1136/bmj.i6460 [published Online First: 2017/01/07]
2. 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: https://doi.org/10.1002/sim.6152
3. Wolbers M, Koller MT, Witteman JC, et al. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology (Cambridge, Mass)* 2009;20(4):555-61. doi: 10.1097/EDE.0b013e3181a39056 [published Online First: 2009/04/16]
4. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30(10):1105-17. doi: 10.1002/sim.4154 [published Online First: 2011/01/13]
5. Wolbers M, Blanche P, Koller MT, et al. Concordance for prognostic models with competing risks. *Biostatistics* 2014;15(3):526-39. doi: 10.1093/biostatistics/kxt059 [published Online First: 2014/02/05]
6. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23(5):723-48. doi: 10.1002/sim.1621 [published Online First: 2004/02/26]
7. Teece L. Investigating the presence and impact of competing events on prognostic model research. Keele University, 2019.
8. 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]