Research Letter format

Title page:

Title: A note on performance metrics for the kidney failure risk equation

Authors:

Oskar Ålund[1], Robert Unwin[2], Benjamin Challis[2], Philip A Kalra[3], Maarten W Taal[4], David C Wheeler[5], Simon DS Fraser[6], Paul Cockwell[7], Magnus Söderberg[1]

Affiliations:

[1]AstraZeneca, Gothenburg, Sweden

[2]AstraZeneca, Cambridge, UK

[3]Dept of Renal Medicine, Salford Royal Hospital and University of Manchester, UK

[4]Centre for Kidney Research and Innovation, University of Nottingham

[5]Department of Renal Medicine, University College London

[6]School of Primary Care, Population Science and Medical Education, Faculty of Medicine, University of Southampton

[7]Department of Renal Medicine, Queen Elizabeth Hospital, University Hospitals of Birmingham, Birmingham, UK

Correspondence to:

Oskar Ålund

Email: oskar.alund@astrazeneca.com

Word count (maximum 1200 words)**: 1265**

Number of figures, tables, boxes, references (maximum 15 references)**:** 2 tables, 1 figure, 15 references.

Page 2: (structured only by paragraphing without subheadings, maximum 1200 words)

The kidney failure risk equation[1] (KFRE) is based on a Cox model first published in 2011 to quantify the risk of initiating kidney replacement therapy (KRT) over 2 and 5 years in patients with stage 3-5 CKD. It is commonly described as an accurate model[1-4] achieving excellent discrimination[5]. These descriptions are grounded in several evaluation metrics, most prominently the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUROC), and Harrell's C-index. However, these metrics exhibit limitations in conveying model performance in the context of highly censored populations with low event rate. In this letter we discuss the shortcomings of the aforementioned evaluation metrics and argue for a greater emphasis on more clinically relevant metrics such as the positive and negative predictive values (PPV and NPV).

Using data from the National Unified Renal Translational Research Enterprise-Chronic Kidney Disease (NURTuRE-CKD) cohort we computed 4-variable KFRE scores for all participants with eGFR less than 60 mL/min/1.73m^2 and a UACR result available at baseline (N=2444, 60% male). Baseline statistics are shown in Table 1. Detailed baseline characteristics of the NURTuRE CKD cohort have been published previously[6].

	mean	**std**	**min**	**25%**	**50%**	**75%**
Age (years)	63.6	14.1	18.0	55.0	66.0	74.0
eGFR (mL/min/1.73m^2)	33.2	12.0	6.2	23.2	32.2	42.4
UACR (mg/g)	787.5	2333.3	1.1	33.4	203.8	897.4
KFRE 2yr risk (%)	8.6	13.7	0.0	0.8	2.8	9.8
KFRE 5yr risk (%)	22.4	26.4	0.1	3.1	10.6	32.9
KRT at 2 years (%)	6.0					
KRT at 5 years (%)	20.8					

Table 1: NURTuRE baseline statistics. The rows labelled „KRT at 2 (5) years" indicate the proportion of participants that had initiated KRT within 2 (5) years.

The mean time between the baseline measurements and the most recently compiled records was 4.4 years. For all 2444 participants, more than 2 years had elapsed between the baseline measurements and the most recently compiled outcome records. For 250 participants, more than 5 years had elapsed. Discrimination at 2 and 5 years was evaluated in terms of AUROC, average PPV, and average NPV. Overall performance was evaluated in terms of Harrell's C-index, both on the full cohort and on the subset of participants who had initiated KRT within the observation period (we call this the *uncensored C-index*). The results are shown in Table 2 and Figure 1.

Metric	**Value**
2-year AUROC	0.911 (95% CI 0.885–0.937)
2-year Average NPV	0.991 (95% CI 0.987–0.995)
2-year Average PPV	0.500 (95% CI 0.412–0.589)
5-year AUROC	0.837 (95% CI 0.778–0.896)
5-year Average NPV	0.947 (95% CI 0.923–0.971)
5-year Average PPV	0.591 (95% CI 0.451–0.732)
Harrell's C-index	0.867 (95% CI 0.850–0.884)

Uncensored C-index	0.674 (95% CI 0.645–0.704)

Table 2: KFRE evaluation metrics. 2-year metrics calculated with respect to all 2444 participants. 5-year metrics calculated with respect to the 250 participants where more than 5 years elapsed since the baseline measurements.

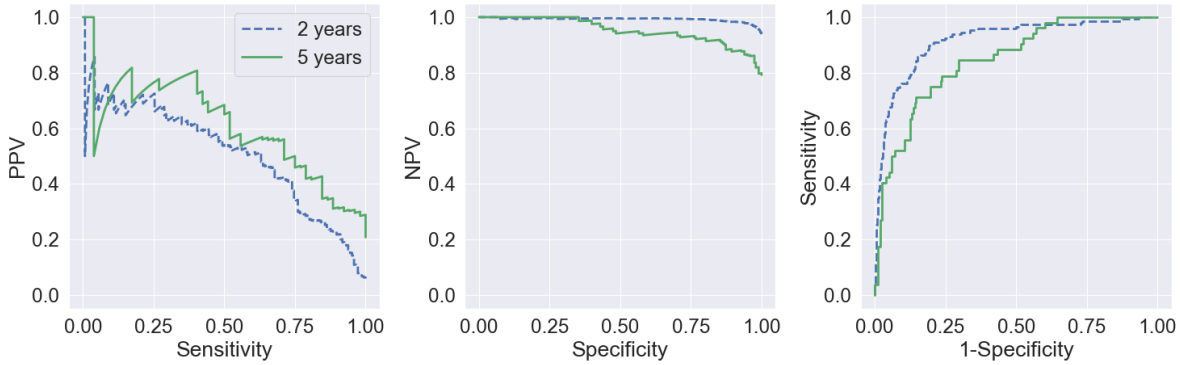

Figure 1: KFRE discrimination curves at 2 and 5 years.

Our findings are in agreement with prior validation studies in terms of the ROC and Harrell's C-index. The NPV was excellent for both 2- and 5-year discrimination. In contrast, however, the PPV was low for both 2- and 5-year discrimination (average PPV 0.50 and 0.59 respectively). Similarly, the uncensored C-index was significantly lower than Harrell's C-index.

It could be argued that PPV, NPV, and the uncensored C-index are more appropriate metrics of clinical utility than the ROC and Harrell's C-index computed on the full cohort. The KFRE—given the patient's age, sex, eGFR, and UACR—produces a risk score quantifying the risk of initiating KRT. Thus, by selecting a risk threshold, we can study the KFRE's ability to differentiate between patients that do and do not initiate KRT. In other words, if we suppose that any patient whose KFRE score exceeds the risk threshold will initiate KRT within a particular timeframe, then to what extent do we correctly differentiate between the two groups? The answer to this question can be captured in four relevant probabilities:

- **Sensitivity:** The probability that a patient will score above the risk threshold given that the patient will initiate KRT.
- **Specificity:** The probability that a patient will score below the risk threshold given that the patient will not initiate KRT.
- **PPV:** The probability that a patient will initiate KRT given that the patient scores above the risk threshold.
- **NPV:** The probability that a patient will not initiate KRT given that the patient scores below the risk threshold.

Note that sensitivity and specificity are not clinically useful probabilities because they are conditioned on what one is trying to assess (initiation of KRT) rather than on what one observes (the risk score). The ROC curve, by definition, is sensitivity and specificity as a function of the risk threshold, and consequently cannot be used alone to argue for or against clinical application. In contrast, PPV and NPV *are* clinically useful probabilities, because they

are conditioned on what one observes (the risk score). Furthermore, by Bayes' theorem, PPV and NPV are functions of sensitivity, specificity, and *prevalence*. More precisely we have

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specificity) \times (1 - prevalence)},$$

and

$$NPV = \frac{specificity \times (1 - prevalence)}{specificity \times (1 - prevalence) + (1 - sensitivity) \times prevalence}.$$

PPV decreases with decreased prevalence, meaning that for fixed sensitivity and specificity, lower prevalence implies lower PPV. In an average population of stage 3-5 CKD patients, the number of patients who initiate KRT within two to five years is low. Therefore, high AUROC does not imply high PPV. Note that average PPV can be artificially increased by considering a higher prevalence population (PPV tends to prevalence as the discrimination threshold tends to -∞). For example, considering only patients with eGFR < 20 mL/min/1.73m^2, average PPV increased to 0.60 and 0.66 for 2- and 5-year discrimination respectively. All evaluation metrics for this subgroup can be found in the online supplementary data.

Harrell's C-index estimates the probability that given two CKD patients, the one with the higher KFRE score will initiate KRT first. It is obtained by forming *comparable* participant pairs and computing the proportion of pairs that the KFRE sorts correctly. The word "comparable" is important and severely limits the usefulness of Harrell's C-index as a performance measure of the KFRE. Two participants are comparable if it can be determined which participant initiated KRT first. In the NURTuRE cohort, and indeed in most CKD cohorts, most participants did not initiate KRT during the observation period. This means that most of the contribution to Harrell's C-index comes from comparing participants that initiated KRT within the observation period to participants that did not. In other words, a high-risk group is being compared to a low-risk group, which causes Harrell's C-index to overestimate the target probability. A more informative statistic is obtained by computing Harrell's C-index on the subset of participants that initiated KRT within the study observation period – our uncensored C-index, which produced a much lower value.

Systematic reviews[7, 8] have correctly pointed out the need for more comprehensive reporting of prediction model performance. We repeat and emphasize this concern. For clinical application, it is essential to be able to clearly quantify confidence in a particular risk prediction in the form of PPV and NPV in various subpopulations. This aspect is absent from several other risk models explored in the literature[9-15].

We conclude that the KFRE is an important tool for epidemiological use. The excellent NPV should reassure clinicians that a patient with low KFRE risk score does not require referral to a nephrologist. On the other hand, we have identified that the KFRE gives a risk estimate with a low PPV. A high predicted risk should therefore be interpreted with some caution and placed in a wider clinical context to guide patient management and follow-up.

It should be noted that commencement of KRT is not a precise outcome; patients with ESKD may choose a conservative course, whereas those who opt for KRT commence this at variable GFR. Unfortunately, a high proportion of patients with advanced CKD die prior to KRT initiation. These aspects are not well captured by the KFRE and likely contribute to the low PPV. Our analysis suggests that linear risk scores based on age, sex, eGFR, and UACR cannot predict the risk of initiating KRT with high PPV in populations with relatively low overall event rate and that alternative approaches to risk prediction should be explored to achieve a higher PPV.

Online supplementary data (optional, unrestricted, a single PDF file with a title page providing the title and a page index)

Acknowledgements:

Funding:

This study is supported by funding provided by several commercial partners (AbbVie, AstraZeneca, Evotec International GmbH, Travere Therapeutics, UCB Biopharma) to Kidney Research UK through a formal collaboration agreement and awarded to the academic investigators as a research grant. The views expressed in this publication are those of the authors and not necessarily those of the funders or sponsor.

Authors' contributions:

Oskar Ålund: Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, Visualization

Robert Unwin: Conceptualization, Writing – Review & Editing

Benjamin Challis: Writing – Review & Editing

Philip A Kalra: Writing – Review & Editing, Data Curation

Maarten W Taal: Writing – Review & Editing, Data Curation

David C Wheeler: Writing – Review & Editing, Data Curation

Simon DS Fraser: Writing – Review & Editing, Data Curation

Paul Cockwell: Writing – Review & Editing, Data Curation

Magnus Söderberg: Conceptualization, Writing – Review & Editing, Supervision, Project administration

Conflict of interest statement:

none declared

References (maximum 15)**:**

1. Tangri N, Stevens LA, Griffith J*, et al.* A predictive model for progression of chronic kidney disease to kidney failure. Jama 2011;305(15):1553-1559
2. Major RW, Shepherd D, Medcalf JF*, et al.* The kidney failure risk equation for prediction of end stage renal disease in UK primary care: an external validation and clinical impact projection cohort study. PLoS medicine 2019;16(11):e1002955
3. Peeters MJ, van Zuilen AD, van den Brand JA*, et al.* Validation of the kidney failure risk equation in European CKD patients. Nephrology Dialysis Transplantation 2013;28(7):1773-1779
4. Whitlock RH, Chartier M, Komenda P*, et al.* Validation of the kidney failure risk equation in Manitoba. Canadian journal of kidney health and disease 2017;4:2054358117705372
5. 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):164-174
6. Taal MW, Lucas B, Roderick P*, et al.* Associations with age and glomerular filtration rate in a referred population with chronic kidney disease: methods and baseline data from a UK multicentre cohort study (NURTuRE-CKD). Nephrology Dialysis Transplantation 2023
7. Collins GS, Omar O, Shanyinde M*, et al.* A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of clinical epidemiology 2013;66(3):268-277
8. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS medicine 2012;9(11):e1001344
9. Ferguson T, Ravani P, Sood MM*, et al.* Development and external validation of a machine learning model for progression of CKD. Kidney International Reports 2022;7(8):1772-1781
10. Schena FP, Anelli VW, Trotta J*, et al.* Development and testing of an artificial intelligence tool for predicting end-stage kidney disease in patients with immunoglobulin A nephropathy. Kidney international 2021;99(5):1179-1188
11. Zacharias HU, Altenbuchinger M, Schultheiss UT*, et al.* A predictive model for progression of CKD to kidney failure based on routine laboratory tests. American Journal of Kidney Diseases 2022;79(2):217-230. e211
12. Lin Y, Shao H, Fonseca V*, et al.* A prediction model on incident ESKD among individuals with T2D and CKD. Journal of Diabetes and its Complications 2023;37(4):108450
13. Goldfarb-Rumyantzev AS, Gautam S, Dong N*, et al.* Prediction model and risk stratification tool for survival in patients with CKD. Kidney international reports 2018;3(2):417-425
14. Xiao J, Ding R, Xu X*, et al.* Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine 2019;17(1):1-13
15. Jardine MJ, Hata J, Woodward M*, et al.* Prediction of kidney-related outcomes in patients with type 2 diabetes. American journal of kidney diseases 2012;60(5):770-778