BMC Nephrology

RESEARCH

Open Access

# Comparative performance of risk prediction models for kidney disease: an external validation using 0.5 million UK Biobank participants

Yikun Zhang[1], Chun Hin Chan[1], Hin Lai Ivan Lam[1], David Bishai[1], Philip Clarke[2], Sydney C.W. Tang[3] and Jianchao Quan[1*]

## Abstract

**Objective** To compare the external validation performance of existing kidney disease risk prediction models for the general population, individuals with type 2 diabetes, and across various subgroups.

**Materials and methods** We identified and compared 16 risk prediction models for chronic kidney disease (CKD) or kidney failure from 3 recent systematic reviews (7 models for the whole population, 9 models specific for type 2 diabetes). We analysed 497,896 adults (age 38–73) in the UK Biobank data; of which 4·7% ($n = 23,298$) had type 2 diabetes. Models were evaluated by discrimination and calibration performance with subgroup analyses by age, sex, ethnicity and pre-existing hypertension.

**Results** During a total follow-up of 5·95 million person-years (median: 12·2 years; IQR: 1·4), predictive models for people without diabetes exhibited fair-to-excellent discrimination performance (c-indices: 0·695-0·806) but severely overpredicted risk. The O'Seaghdha model demonstrated the best overall performance for discrimination (c-index: 0·806 [0·806-0·807]) and calibration (slope: 0·69, intercept: -0·011; Brier score: 0·03 [0·02,0·04]). Models including medications for diabetes showed superior performance. Discriminative performance was poorer for people with diabetes or hypertension. Severe miscalibration occurred for many models.

**Conclusion** Most models demonstrated fair to excellent discrimination for CKD and good to excellent discrimination for kidney failure. Calibration performance was frequently suboptimal; most models substantially overpredicted CKD risk while underpredicting kidney failure risk, indicating that recalibration is warranted prior to clinical application. Model performance in individuals with diabetes or hypertension was poorer. Future CKD risk prediction model development should incorporate diabetes medication use to enhance discriminative capability.

**Keywords** Chronic kidney disease, Kidney failure, Type 2 diabetes, Hypertension, External validation, Prediction model

*Correspondence:
Jianchao Quan
jquan@hku.hk
[1]School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Patrick Manson Building, 7 Sassoon Road, Pokfulam, Hong Kong, China
[2]Health Economics Research Centre & REAL Demand Unit, Nuffield Department of Population Health, University of Oxford, Oxford, Nuffield, UK
[3]Division of Nephrology, Department of Medicine, School of Clinical Medicine, The University of Hong Kong, Queen Mary Hospital, Hong Kong, China

## Introduction

Chronic kidney disease (CKD) is a leading cause of mortality andmorbidity and has been forecasted to becomethe fifth leading cause ofworldwide deaths by 2050. Theburden of CKD is projected to increase dueto population aging and the rising prevalence of diabetes and hypertension [1]. Diabetesis the leading cause of CKD and the prevalence of both conditions is on the rise [2]. The occurrence of CKD in individuals with diabetesis associated with increased risks of kidneyfailure, cardiovascular disease and premature mortality [3]. Current recommendations highlight the importance of earlyscreening for CKD in people with type 2 diabetes to enable treatments that improve patient outcomes and minimize healthcare costs [4]. As CKD is a complex condition with multiple risk factors, validated risk prediction models are essential for accurately identifying at-risk individuals and supporting targeted management, especially in outpatient and primary care settings.

Risk prediction models may significantly impact future CKD management, but none have been standardized for consistent clinical use despite widespread development across populations. A recent systematic review identified 36 risk models for healthy populations and 12 risk models for type 2 diabetes highlighting the broad scope and diversity of existing models [5]. While another systematic review and external validation of 21 prediction models in a Dutch diabetes cohort found good discrimination and validation [6], a meta-analysis of risk models for people with type 2 diabetes noted a high risk of bias in their validation [7]. These studies underscore the need for comparative assessment and external validation of risk models across diverse populations to determine their generalizability beyond specific disease cohorts. To address these uncertainties, we externally validated publicly available risk prediction models for the onset of CKD or kidney failure among individuals without pre-existing kidney disease, using a large population-based cohort of 0.5 million people from the UK Biobank. Given the considerable variation in CKD prevalence between individuals with and without diabetes, we evaluated model performance separately in these subgroups [8].

## Materials and methods

### Model selection

We identified risk prediction models for CKD and kidney failure in both general population and people with type 2 diabetes from three recently published systematic reviews [1–3]. A total of 20 models was identified for general population, of which 13 models were excluded (Table 1). We validated the seven remaining risk models for CKD (Chien, Nelson, Kshirsagar's simplified categorical model, Kshirsagar's best-fitting categorical model, O'Seaghdha, Saranburut, and Umesawa) [9–14]. We

**Table 1** Risk prediction models included in this study

| Model/Author | Location | Publication year | Study period of derivation cohort | Predicted time |
|---|---|---|---|---|
| **General population** | | | | |
| Chien (5) (Clinical model) | Taiwan | 2010 | 2003–2007 | 4 years |
| Kshirsagar (6) (Simplified categorical model) | USA | 2008 | 1987–2003 | 10 years |
| Kshirsagar (6) (Best-fitting categorical model) | USA | 2008 | 1987–2003 | 10 years |
| Nelson (7) | Multi-national (28 countries) | 2019 | 1970–2017 | 5 years |
| O'Seaghdha (8) | USA | 2012 | 1995–2008 | 10 years |
| Saranburut (9) | Thailand | 2017 | 1985–2014 | 10 years |
| Umesawa (10) | Japan | 2017 | 1993–2006 | 10 years |
| ***People with type 2 diabetes*** | | | | |
| Dunkler (11) (Laboratory model) | Multi-national (40 countries) | 2015 | 2003–2011 | 5.5 years |
| Jardine (12) | Multi-national (20 countries) | 2012 | 2001–2008 | 5 years |
| Low (13) | Singapore | 2017 | 2002–2014 | 6 years |
| UKPDS OM2 (14) | UK | 2013 | 1997–2007 | 5 years |
| ZODIAC-36 (15) (Cox regression model) | Netherlands | 2015 | 1998–2009 | 10 years |
| ZODIAC-36 (15) (Competing risk model) | Netherlands | 2015 | 1998–2009 | 10 years |
| RECODe (16) | USA | 2017 | 2001–2009 | 10 years |
| Wan (17) | China | 2017 | 2010–2015 | 5 years |
| Elley (18) | New Zealand | 2013 | 2000–2010 | 5 years |

identified a further 31 models specifically for individuals with type 2 diabetes, of which 9 models were included in the analysis (Dunkler, Low, ZODIAC-36's cox regression model, ZODIAC-36's competing risk model, UKPDS OM2, Jardine, RECODe, Wan, and Elley) with 13 unique equations for CKD and kidney failure [15–22]. The remaining 22 models were excluded (detailed in Figure S1). Models were excluded if the predictive equations were not publicly available or if the required predictor variables were unavailable in the dataset. Models developed for patients with existing CKD or those predicting composite outcomes were also excluded to ensure consistency with the study population and outcome definitions. Additionally, models with a prediction horizon of less than three years were excluded, as short-term risk predictions are considered less clinically actionable for the prevention and long-term management of CKD, a progressive condition. A literature search from December 2021 to January 2025 identified no additional CKD risk prediction models beyond those found in the reviews.

#### Validation data

The validation dataset was derived from the UK Biobank, a population-based cohort of 500,000 participants recruited from 2006 to 2010 in 22 assessment centres across the UK [23]. Hospital inpatient data, used in this study, were linked to NHS England, the Information and Statistics Division of Scotland, and the Secure Anonymised Information Linkage system in Wales [24]. Our right-censored end-dates were 26 April 2021 and 5 May 2021 for cohorts with and without type 2 diabetes respectively. Data on UK Biobank participants include demographics, biomarker values, medication usage, disease diagnoses and healthcare service utilisation. The eGFR was calculated using the 2021 CKD-EPI equation. Diabetes status was coded as a binary variable. Other disease diagnoses, include CKD and kidney failure, were coded using the International Classification of Diseases, Tenth Revision (ICD-10). Our analysis separated the validation cohort into people with and without type 2 diabetes. The validation cohort without diabetes had 474,598 individuals after excluding people with diabetes and pre-existing CKD or kidney failure. For the type 2 diabetes validation cohort, we identified people with recorded diabetes (excluding gestational diabetes and type 1 diabetes) and excluded those with pre-existing CKD or kidney failure, resulting in a total of 23,298 individuals.

#### Outcome

The outcomes of the validated models were CKD and kidney failure. The endpoint was the first occurrence of recorded diagnosis of CKD or kidney failure. Diagnosis records were coded using ICD-10. We tailored the outcome definitions to match each specific risk model. For models that did not provide specific ICD codes, we defined the CKD as abnormalities of kidney structure or function, as indicated by either estimated eGFR or albuminuria; and specified the CKD (R80, N181, N182, N183, N184, N189, N289, E112, E142, N083, N180, N185, N19, Z940, Z992, Y841) and kidney failure (N180, N185, N19, Z940, Z992, Y841) codes. To enhanced comparability and consistency across the various risk prediction models, we also present results using a uniform outcome definition aligned with these ICD codes. Detailed definitions used for each model are in Table S1.

#### Statistical analysis

We truncated the UK Biobank validation data to match the fixed prediction timespan for each model. For UKPDS OM2, we selected a 5-year time frame, consistent with most type 2 diabetes models. Predicted risks were calculated using the UKPDS OM2 Simulation Software [25]. Sex, smoking status, medication usage, and disease history were coded as binary values with numerical unit conversions where necessary.

Model performance was assessed using two key metrics: discrimination and calibration. Discrimination evaluates the model's ability to differentiate between positive and negative outcomes and was measured using Uno's concordance index (c-index), which accounts for the presence of competing risks. The c-index ranges from 0 to 1, with values closer to 1 indicating better discriminative performance. The c-index calculates the area under the time-dependent receiver operating characteristics curve [26]. To enhance the robustness of these results, traditional ROC curves and the area under the ROC curve (AUC) were also evaluated. Performance interpretation was based on the following thresholds: average (0.50–0.60), fair (0.60–0.70), good (0.70 – 0.80), and excellent ($\geq$ 0.8) [27]. The 95% confidence intervals were calculated by bootstrapping 100 replications. Calibration evaluates the agreement between the observed and predicted values, was assessed with calibration plots, targeting an ideal slope of 1 and intercept of 0. Mild-to-moderate miscalibration was defined as a slope between 0.7 and 1.3. Predictive accuracy was quantified using the Brier score—the mean squared difference between predicted probabilities and observed outcomes—ranging from 0 to 1, with lower scores indicating greater accuracy [28]. Subgroup analyses were performed by sex, age, ethnicity, and hypertension status. Additionally, decision curve analysis (DCA) was employed to assess clinical utility. DCA estimates the net benefit, a weighted measure of true positives against false positives, to determine whether model-based clinical decision-making offers added value over default strategies of treating all or no patients. For each model, net benefit was calculated and

plotted across a range of decision thresholds, enabling direct comparison of their expected clinical usefulness.

We applied multiple imputation to predictors with < 20% missing values. This threshold for missingness is supported as robust by recent literature [29]. For predictors with > 20% missing values, we imputed the sample means of the derivation dataset used for each model. The results were combined using Rubin's Rules to account for variability across the imputed datasets [30]. To assess the impact of missing data, we performed sensitivity analyses using both complete case analysis (CCA) and, specifically for urinary albumin, multiple imputation as an alternative approach. All imputations were conducted using the *Hmisc* [31] and *mice* [32] packages in R statistical software.

## Results

We evaluated a total of 16 models (seven developed for general population and nine models specifically for individuals with type 2 diabetes). Models were published from 2008 to 2019 and were developed across Western countries, Asia, or multi-national populations. These models estimated the risk of CKD or kidney failure over prediction horizons ranging from 4 to 10 years from baseline. The study period for the derivation cohort spanned from the 1970s to the 2010s (Table 1).

Table 2 and Table S2 show the baseline characteristics of validation dataset of 474,598 participants without diabetes and 23,298 participants with type 2 diabetes in the UK Biobank. The mean age at diagnosis for diabetes was 52.6 years (SD = 9.7), and the mean duration of diabetes was 7.9 years (SD = 7.5). Table S3 presents the differences in characteristics between subjects with and without missing urinary albumin data. All characteristics were similar except for higher urinary creatinine among the included participants compared to those excluded from the complete case analysis (median: 12,287 mmol/L vs. 5,936 mmol/L for individuals without diabetes, 11,281 mmol/L vs. 6,519 mmol/L for those with diabetes).

Most models included common predictors such as age, sex, BMI, blood pressure, and eGFR (Table S4). Table S5 shows the number of events and event rate for each outcome. In the cohort without diabetes, there were 15,670 incidences of CKD (prevalence 3.3%) over 5,684,699 person-years of follow-up (mean = 12.0 years, median = 12.2; IQR: 1.4). In the cohort with type 2 diabetes, there were 3,598 incidences of CKD (prevalence 15.4%) and 561 incidences of kidney failure (prevalence 2.4%) recorded over 265,852 person-years of follow-up (mean = 11.4 years, median = 11.9 years; IQR: 1.6).

### Validation performance

The risk prediction models for the general population without diabetes exhibited good to excellent discriminative performance for CKD with c-indices between 0.69 and 0.81 (Table 3). The O'Seaghdha model exhibited highest discrimination with c-indices of 0.81. This finding was supported by consistent results from the ROC curve and the AUC. (Figure S2) In subgroup analysis (Fig. 1), the models demonstrated slightly better discrimination for younger individuals compared to older ones (0.627–0.766 vs. 0.589–0.765). The c-index for non-white ethnicities in Kshirsagar's best-fitting categorical model is not presented in the figure due to insufficient data for subgroup analysis. The performance among people with hypertension was lower than those without hypertension (0.586–0.770 vs. 0.658–0.793). There was no discernible pattern by sex nor ethnicity.

All models for people without diabetes tended to overpredict the risk of CKD (slope range: 0.13–0.69; intercept range: -0.020–0.002) (Fig. 2). The O'Seaghdha model demonstrated the best calibration performance with a slope of 0.69, an intercept of -0.011. This model consistently exhibited the best calibration performance across all subgroups, except in male subgroup (Figure S3). No apparent differences in calibration performance were observed between the subgroups. Brier scores across the evaluated models ranged from 0.03 to 0.18, with the O'Seaghdha model achieving the lowest score of 0.03, indicating the highest overall predictive accuracy (Table S6). Although the Nelson and Umesawa models demonstrated high discrimination, with C-indices exceeding 0.8; both models tended to overpredict risk, yielding Brier scores of 0.17 and 0.11, respectively.

DCA demonstrated that clinical utility was limited for all seven models when applied to individuals without diabetes. (Figure S4) At very low threshold probabilities (< 3%), most model yielded modest positive net benefits exceeding both default strategies. The Seaghdha model provided a positive net benefit at the higher threshold of 8%. However, across the broader clinically relevant threshold range (5–30%), all models approximated or fell below the "Treat None" reference, indicating that model-guided decision-making conferred no incremental benefit over not intervening.

### Population with type 2 diabetes

In individuals with type 2 diabetes, model discrimination for CKD ranged from 0.603 to 0.758 and for kidney failure from 0.704 to 0.880 (Table 3). Using a uniform outcome definition yielded results similar to the individualized model-specified definitions. An exception was the ZODIAC-36 model -designed for early-stage CKD—demonstrated reduced performance in this standardized framework. The Nelson, Umesawa and O'Seaghdha models were performing well in predicting CKD, (c-indices: 0.758, 0.753, and 0.737) and Elley model for predicting kidney failure (c-index: 0.880) (Table 3). The results from

**Table 2** Baseline characteristics of UK biobank validation dataset

| Variables | *n* (%) or mean (± SD) | |
|---|---|---|
| | **Cohort without diabetes (*n* = 474,598)** | **Cohort with type 2 diabetes (*n* = 23,298)** |
| Demographics | | |
| Age (years) | 56.9 (± 8.1) | 60.5 (± 6.9) |
| 40–49 | 107,636 (22.7) | 2,087 (9.0) |
| 50–59 | 156,114 (32.9) | 6,479 (27.8) |
| 60–69 | 201,451 (42.4) | 13,920 (59.7) |
| 70 and above | 9,393 (2.0) | 815 (3.5) |
| Ethnicity | | |
| White | 448,387 (94.5) | 20,361 (87.4) |
| Other | 26,211 (5.5) | 2,937 (12.6) |
| Sex | | |
| Male | 212,201 (44.7) | 14,744 (63.3) |
| Female | 262,396 (55.3) | 8,554 (36.7) |
| **Clinical features** | | |
| Smoking status | | |
| Never | 260,739 (54.9) | 10,368 (44.5) |
| Former smoker | 161,238 (34.0) | 10,209 (43.8) |
| Current smoker | 49,868 (10.5) | 2,570 (11.0) |
| Diabetes | | |
| Age at diagnosis of diabetes (year) | - | 52.6 (± 9.7) |
| Duration of diabetes (years) | - | 7.9 (± 7.5) |
| BMI (kg/m$^2$) | 27.2 (± 4.6) | 31.5 (± 5.8) |
| Diastolic blood pressure (mmHg) | 82.3 (± 10.7) | 81.5 (± 10.3) |
| Systolic blood pressure (mmHg) | 139.5 (± 19.7) | 143.6 (± 18.5) |
| **Biomarkers** | | |
| Urate (umol/L) | 308 (± 79.7) | 333.6 (± 85.3) |
| Hba1c (mmol/mol)* | 35.4 (5.0) | 51.7 (16.7) |
| Cholesterol (mmol/L) | 5.8 (± 1.1) | 4.5 (± 1.0) |
| LDL (mmol/L) | 3.6 (± 0.9) | 2.7 (± 0.7) |
| HDL (mmol/L)* | 1.4 (0.5) | 1.1 (0.4) |
| Triglycerides (mmol/L)* | 1.5 (1.1) | 1.9 (1.3) |
| White blood cell (10^9 cells/L)* | 6.6 (2.2) | 7.4 (2.5) |
| Creatinine (mmol/L)* | 70.7 (17.7) | 70.7 (17.7) |
| Urinary Creatinine (mmol /L)* | 7477 (7682.0) | 8,832 (7437) |
| Albumin (g/L) | 45.2 (± 2.6) | 45.0 (± 2.9) |
| Albumin in urine (mg/L)* | 11.2 (10.1) | 14.1 (25.4) |
| eGFR (ml/min/1.73 m$^2$)* | 85.8 (20.3) | 87.8 (26.3) |

* median (IQR)

the ROC curves and corresponding AUCs were consistent with the c-index findings. (Figure S5, Figure S6) Most models demonstrated slightly better discrimination for non-white ethnicities than white ethnicities in predicting both CKD (0.623–0.776 vs. 0.596–0.754) and kidney failure (0.721–0.884 vs. 0.663–0.877). Minimal variation was observed between sex, age and hypertension status subgroups. (Figure S7, Figure S8)

In terms of calibration, the O'Seaghdha had the best performance with calibration slopes of 0.94 (intercept – 0.034). Most models tended to overpredict CKD (slope range: 0.22–0.58, intercept range: -0.055-0.045), except for UKPDS OM2 and ZODIAC-36 (Fig. 2); and underpredict kidney failure (slope range: 3.08–5.29, intercept range: -0.081-0.006; Figure S9), except for ZODIAC-36 and the Wan model. The O'Seaghdha model demonstrated good performance in predicting CKD across most subgroups. (Figure S10). All models predicting kidney failure showed severe miscalibration. There was no apparent difference in calibration performance between subgroups (Figure S11). Brier scores for the prediction of CKD ranged from 0.02 to 0.39, while Brier scores for the prediction of kidney failure ranged from 0.01 to 0.10. The overall high accuracy observed for kidney failure predictions, as reflected by lower Brier scores, was likely influenced by the low event rate of kidney failure compared to CKD (Table S6).

**Table 3** Discriminative performance of the prediction models

| Model | C-index [95% CI] | |
|---|---|---|
| | Study-Specific Outcome Definition | Standardized Outcome Definition |
| **People without diabetes** | | |
| ***Chronic Kidney Disease*** | | |
| Chien | 0.719 [0.719–0.720] | 0.719 [0.719–0.720] |
| Kshirsagar (Simplified categorical model) | 0.695 [0.695–0.696] | 0.695 [0.695–0.696] |
| Kshirsagar (Best-fitting categorical model) | 0.708 [0.708–0.709] | 0.708 [0.708–0.709] |
| Nelson | 0.803 [0.803–0.804] | 0.803 [0.803–0.804] |
| O'Seaghdha | 0.806 [0.806–0.807] | 0.806 [0.806–0.807] |
| Saranburut | 0.772 [0.772–0.773] | 0.772 [0.772–0.773] |
| Umesawa | 0.800 [0.799-0.800] | 0.800 [0.799-0.800] |
| **People with type 2 diabetes** | | |
| ***Chronic Kidney Disease*** | | |
| Chien | 0.603 [0.602–0.603] | 0.603 [0.602–0.603] |
| Kshirsagar (Simplified categorical model) | 0.652 [0.652–0.653] | 0.652 [0.652–0.653] |
| Kshirsagar (Best-fitting categorical model) | 0.658 [0.658–0.659] | 0.658 [0.658–0.659] |
| Nelson | 0.758 [0.757–0.759] | 0.758 [0.757–0.759] |
| O'Seaghdha | 0.737 [0.736–0.737] | 0.737 [0.736–0.737] |
| Saranburut | 0.716 [0.715–0.717] | 0.716 [0.715–0.717] |
| Umesawa | 0.753 [0.751–0.754] | 0.753 [0.751–0.754] |
| Dunkler (Laboratory model) | 0.615 [0.612–0.618] | 0.615 [0.612–0.618] |
| Jardine | 0.643 [0.642–0.643] | 0.639 [0.639–0.640] |
| Low | 0.621 [0.621–0.622] | 0.621 [0.621–0.622] |
| *UKPDS OM2 | 0.676 [0.666–0.685] | 0.669 [0.656–0.677] |
| ZODIAC-36 (Cox regression model) | 0.696 [0.695–0.697] | 0.603 [0.603–0.604] |
| ZODIAC-36 (Competing risk model) | 0.699 [0.698-0.700] | 0.610 [0.609–0.610] |
| ***Kidney failure*** | | |
| Elley | 0.880 [0.879–0.881] | 0.811 [0.810–0.813] |
| Jardine | 0.806 [0.805–0.808] | 0.806 [0.805–0.808] |
| RECODe | 0.746 [0.743–0.748] | 0.746 [0.743–0.748] |
| Wan | 0.704 [0.703–0.704] | 0.718 [0.717–0.718] |
| ZODIAC-36 (Cox regression model) | 0.672 [0.672–0.673] | 0.672 [0.672–0.673] |
| ZODIAC-36 (Competing risk model) | 0.682 [0.681–0.682] | 0.682 [0.681–0.682] |

* Validation statistics [95%CI]. The results obtained from the UKPDS OM2 simulation software were validated using one dataset for each outcome

DCA indicated that most models predicting CKD provided meaningful clinical utility for predicting CKD among individuals with type 2 diabetes (Figure S12). In particular, the Saranburut, O'Seaghdha, and Umesawa models yielded sustained positive net benefit exceeding both the "Treat All" and "Treat None" strategies across a broad range of threshold probabilities (up to approximately 25–30%). For kidney failure prediction, DCA revealed generally modest clinical utility across all models, consistent with the low prevalence of this outcome (Figure S13). The Elley and Jardine models demonstrated the most sustained positive net benefit, exceeding the "Treat None" strategy across the threshold probability range of 0 to 30%, suggesting potential clinical value for risk-guided decision-making.

### Sensitivity analysis

Multiple imputation of missing data did not impact the performance of general risk models for with people without diabetes; c-index values remained within the 95% CIs of the CCA. In models for people with type 2 diabetes, imputation resulted in slightly lower c-indices (CCA:

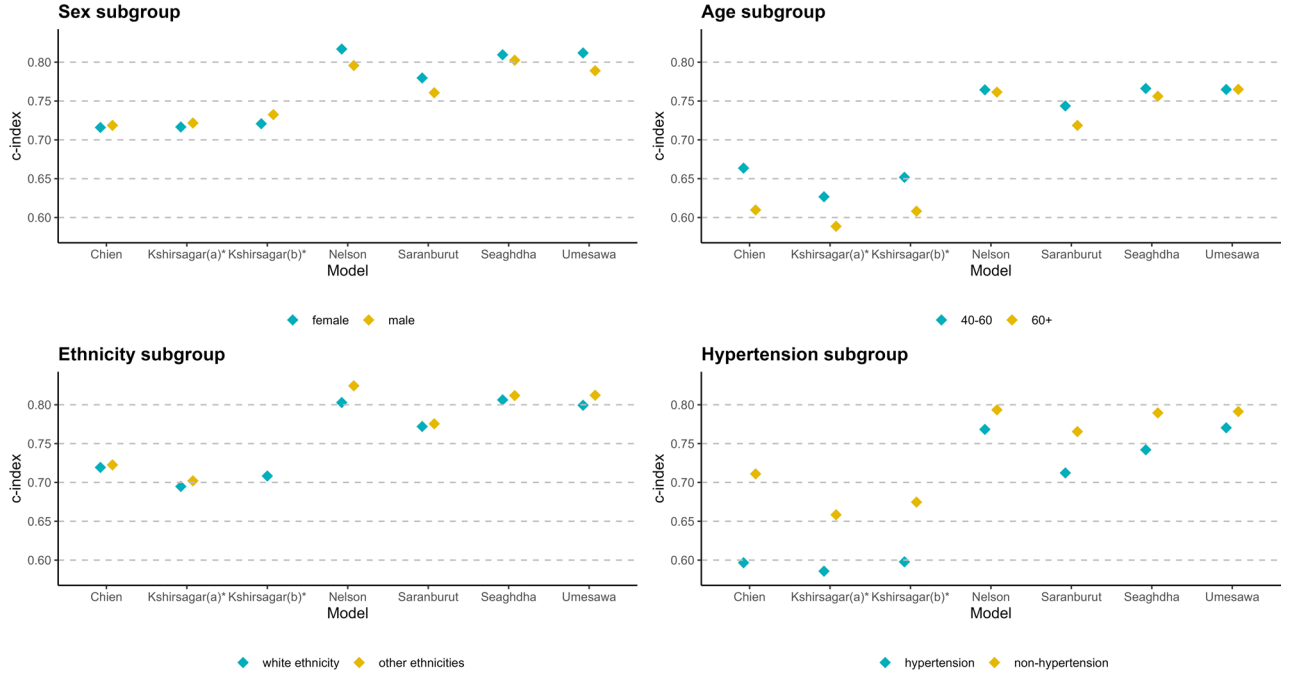


**Fig. 1** Discriminative performance of risk models by subgroup for population without diabetes. * Kshirsagar (**a**): Kshirsagar (simplified categorical model), Kshirsagar (**b**): Kshirsagar (best-fitting categorical model)

0.587–0.768 vs. post-imputation: 0.603–0.758) for CKD prediction, except for the UKPDS OM2 model, where discrimination improved (c-index increased from 0.589 to 0.676 as sample size grew from 8,222 to 21,460). For kidney failure prediction, the range of c-index values were within the 95% CIs of the CCA results, except for the two ZODIAC-36 models (CCA: Cox regression model: 0.746, competing risk model: 0.734 vs. post-imputation: Cox regression model: 0.682, competing risk model: 0.672). Calibration slopes did not improve after imputation. (Table S7) When missing values for urinary albumin was handled using multiple imputation (as opposed to mean imputation), discrimination performance improved for models in individuals without diabetes (mean imputation range: 0.695–0.806; multiple imputation range: 0.719–0.833) and for most kidney failure prediction models in the type 2 diabetes cohort (mean imputation range: 0.672–0.880; multiple imputation range: 0.830–0.961), except for the Wan model. However, for CKD prediction in people with type 2 diabetes, discrimination performance worsened for several models, including Saranburut, Dunkler, and both ZODIAC-36 variants (Cox and competing risk). Calibration slopes did not show any notable improvement following imputation in these analyses. (Table S8)

## Discussion

We externally validated 16 published risk models for predicting CKD and kidney failure: seven models for the general population and nine models specifically for people with type 2 diabetes. Most models showed fair to excellent discrimination for CKD and good to excellent discrimination for kidney failure, though performance was slightly lower among those with diabetes or hypertension. Models incorporating medication use as predictors generally had better discriminatory power. Calibration performance was often suboptimal with most models severely overpredicted CKD and underpredicted kidney failure. The O'Seaghdha model demonstrated the best overall performance with relatively few predictors, suggesting its strong clinical applicability.

The models consistently overpredicted CKD risk in the general population without diabetes with more pronounced overprediction in models with shorter prediction times, e.g. Chien (4 years) and Nelson (5 years). In contrast, models for individuals with type 2 diabetes demonstrated better calibration performance though they still tended to overpredict CKD risk. However, these models generally underestimated the risk of kidney failure; predicting severe kidney complications associated with diabetes remains a challenge that warrants further development. These patterns of miscalibration highlight the necessity of recalibrating risk prediction models prior to clinical application. To address this, we performed intercept-only recalibration and re-assessed calibration performance (Table S9). The results demonstrated that even simple recalibration can significantly enhance calibration performance in most models. Additionally, the ZODIAC-36 model, which initially exhibited low predicted risks and an excessively high calibration slope,

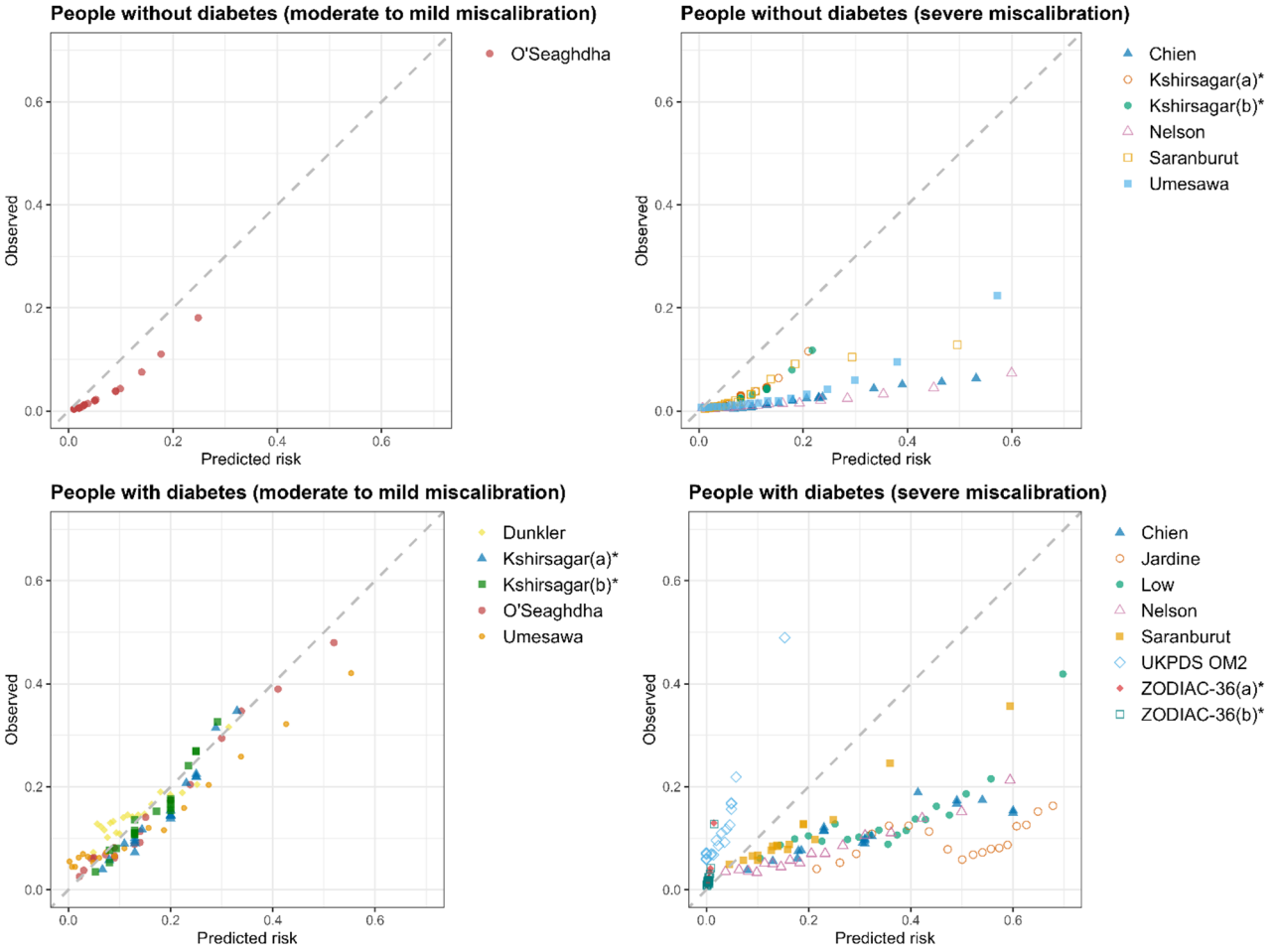


**Fig. 2** Calibration plots for models predicting chronic kidney disease. * ZODIAC-36 (**a**): ZODIAC-36 (cox regression model), ZODIAC-36 (**b**): ZODIAC-36 (competing risk model). Kshirsagar (**a**): Kshirsagar (simplified categorical model), Kshirsagar (**b**): Kshirsagar (best-fitting categorical model)

showed substantial improvement following recalibration. This suggests that the initial miscalibration was predominantly attributable to systematic baseline differences (such as variations in outcome rates) rather than deficiencies in the model's discrimination of individual risk.

The discriminative performance of risk prediction models was lower for people with hypertension in predicting CKD. Risk prediction for people with hypertension may be complicated by the use of antihypertensive medications to protect kidney function. Notably the Umesawa model that includes antihypertensive medication as a predictor exhibited a smaller performance gap between people with and without hypertension. While most models included hypertension as a predictor, they did not consider use of antihypertensive medication. Models showed slightly better discrimination among younger individuals, potentially due to the increased complexity of comorbidities in older age groups. Performance differences between ethnicities are expected given the development data used – some were multi-national with only 5 models developed exclusively from data in western countries. The UK Biobank is predominately white (87.4% and 94.5% of the validation sample with and without diabetes) and variations in sample size and heterogeneity among non-white ethnicities —including mixed, Asian, and Black backgrounds—might also contribute differences in performance.

The risk models analysed in this study demonstrated similar discriminative performance in this independent external validation against the UK Biobank dataset. While six of the risk models for the general population had previous externally validation, most of the models specifically for people with diabetes (six out of nine) had only undergone internal validation. Of the three externally validated models for people with diabetes, Dunkler and Nelson models exhibited weaker performance in this study compared to the external validation results in their original publications, whereas RECODe demonstrated stronger discriminative performance in this independent external validation than reported in the original publication.

Our study possessed several strengths. We validated openly available models identified globally from systematic reviews against a large dataset to ensure objective and robust external validation. We included models for the general population and models specifically for individuals with type 2 diabetes. We independently validated the models using both specific outcome definitions and uniform definitions and conducted robust sensitivity analyses. In addition to evaluating discrimination and calibration, we employed the Brier score to quantify predictive accuracy, providing a comprehensive assessment of model performance. Moreover, the UK Biobank is a large and reliable dataset of over 500,000 participants with a long follow-up time of over ten years. Previous studies have shown that using extensive administrative health records with ICD-10-based definitions can accurately identify patients with kidney failure [33].

Nevertheless, our study was subject to several limitations. Of the 16 models evaluated, 10 lacked specific ICD codes for outcome definitions, limiting their external applicability. The use of ICD codes to identify cases also presented multiple challenges. The use of ICD codes to identify cases also presented multiple challenges. Previously, CKD was defined solely by eGFR [< 60 mL/min/1.73 $m^2$], until the 2013 KDIGO guidelines introduced a combined definition using both eGFR and urinary albumin [34]. Consequently, ICD codes assigned before 2013 may not capture all CKD cases, particularly those in the early stages. For instance, a study from Australia found ICD-10 codes failed to identify 45.9% of CKD cases among patients admitted to general medicine [35]. Such under-documentation in ICD-based case ascertainment may have partially contributed to the overestimation of CKD prevalence observed in our study. Figure S14 illustrates the proportion of participants at baseline, classified by KDIGO based on GFR and urinary albumin levels, after excluding patients with CKD identified using ICD-10 codes. Moreover, participants included in the complete case analysis exhibited higher urinary albumin levels than those excluded, potentially leading to an overestimation of outcomes and affecting the study's generalizability. The elevated urinary creatinine observed among the included participants may be explained by their marginally greater BMI and mean age. Previous studies have shown that urinary creatinine excretion increases significantly with BMI [36]. Furthermore, although we performed intercept-only recalibration and reassessed calibration performance, future research should consider more complex recalibration, particularly for models developed for individuals with type 2 diabetes or hypertension.

Aside from developing new models to predict the risk of CKD or kidney failure, external validation is a crucial step to improve the accuracy and facilitate adoption in clinical practice. Global disparities in the burden and care of CKD have been previously identified, particularly in low-income and lower-middle-income countries [37]. There is a need to validate these CKD prediction models in different regions to help reduce global disparities in the burden and care of CKD.

In conclusion, in this independent comparative assessment of 16 predictive risk models for CKD and kidney failure in a large population-based cohort, some models achieved fair to excellent discrimination performance. The O'Seaghdha model demonstrated the best overall performance in this large UK cohort. The discriminative performance for predicting CKD among individuals with diabetes or hypertension was poorer indicating a need for further development in these high-risk populations. Models including diabetes medication as a predictor demonstrated superior discrimination performance. Recalibration prior to clinical application is needed for most models as they tended to severely overpredict CKD but underpredicted kidney failure.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12882-026-04991-1.

Supplementary Material 1

### Acknowledgements

Preliminary results from this study was previously presented at the ISPOR Real-World Evidence Summit 2025 in Tokyo, Japan (September 2025) and is available in the ISPOR HEOR Presentations Database (link). We acknowledge the opportunity to share our preliminary findings at this conference.

### Author contributions

Yikun Zhang: Methodology, Software, Investigation, Data Curation, Formal analysis, Validation, Visualization, Writing - Original Draft. Chan Chun Hin: Data Curation, Validation, Writing - Original Draft. Lam Hin Lai Ivan: Validation, Writing - Original Draft. David Bishai: Interpretation, Writing - Review & Editing. Philip Clarke: Interpretation, Writing - Review & Editing. Sydney C.W. Tang: Interpretation, Writing - Review & Editing. Jianchao Quan: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition.

### Funding

Research Grants Council of the Hong Kong Special Administrative Region, China (27112518).

### Data availability

The data that support the findings of this study are available from UK Biobank, but restrictions apply to the availability of these data, which were used under licence for the current study and are not publicly available. The data are available upon reasonable request and with permission from UK Biobank. The analysis codes are available from the authors upon reasonable request.

## Declarations

### Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the Institutional Review Board of the University of Hong Kong (UW 21–414). All data utilized in this study were obtained from the UK Biobank Resource. The UK Biobank has received Research Tissue Bank approval from the Northwest Multi-centre Research Ethics Committee (ID: 16/

NW/0274). All participants provided informed consent electronically at the time of recruitment.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 28 January 2026 / Accepted: 10 April 2026
Published online: 15 April 2026

**References**

1. Vollset SE, Ababneh HS, Abate YH, Abbafati C, Abbasgholizadeh R, Abbasian M, et al. Burden of disease scenarios for 204 countries and territories, 2022–2050: a forecasting analysis for the Global Burden of Disease Study 2021. Lancet. 2024;403(10440):2204–56.
2. Thomas MC, Cooper ME, Zimmet P. Changing epidemiology of type 2 diabetes mellitus and associated chronic kidney disease. Nat Rev Nephrol. 2016;12(2):73–81.
3. Afkarian M, Zelnick LR, Hall YN, Heagerty PJ, Tuttle K, Weiss NS, et al. Clinical Manifestations of Kidney Disease Among US Adults With Diabetes, 1988–2014. JAMA. 2016;316(6):602–10.
4. Skolnik NS, Style AJ. Importance of Early Screening and Diagnosis of Chronic Kidney Disease in Patients with Type 2 Diabetes. Diabetes Ther. 2021;12(6):1613–30.
5. González-Rocha A, Colli VA, Denova-Gutiérrez E. Risk Prediction Score for Chronic Kidney Disease in Healthy Adults and Adults With Type 2 Diabetes: Systematic Review. Prev Chronic Dis. 2023;20:E30.
6. Slieker RC, van der Heijden AAWA, Siddiqui MK, Langendoen-Gort M, Nijpels G, Herings R, et al. Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and external validation study. BMJ. 2021;374:n2134.
7. Buchan TA, Malik A, Chan C, Chambers J, Suk Y, Zhu JW, et al. Predictive models for cardiovascular and kidney outcomes in patients with type 2 diabetes: systematic review and meta-analyses. Heart. 2021;107(24):1962–73.
8. Plantinga LC, Crews DC, Coresh J, Miller ERI, Saran R, Yee J, et al. Prevalence of Chronic Kidney Disease in US Adults with Undiagnosed Diabetes or Prediabetes. Clin J Am Soc Nephrol. 2010;5(4):673.
9. Chien KL, Lin HJ, Lee BC, Hsu HC, Lee YT, Chen MF. A Prediction Model for the Risk of Incident Chronic Kidney Disease. Am J Med. 2010;123(9):836–e8462.
10. Kshirsagar AV, Bang H, Bomback AS, Vupputuri S, Shoham DA, Kern LM, et al. A Simple Algorithm to Predict Incident Kidney Disease. Arch Intern Med. 2008;168(22):2466–73.
11. Nelson RG, Grams ME, Ballew SH, Sang Y, Azizi F, Chadban SJ, et al. Development of Risk Prediction Equations for Incident Chronic Kidney Disease. JAMA. 2019;322(21):2104–14.
12. O'Seaghdha CM, Lyass A, Massaro JM, Meigs JB, Coresh J, D'Agostino RB, et al. A Risk Score for Chronic Kidney Disease in the General Population. Am J Med. 2012;125(3):270–7.
13. Saranburut K, Vathesatogkit P, Thongmung N, Chittamma A, Vanavanan S, Tangstheanphan T, et al. Risk scores to predict decreased glomerular filtration rate at 10 years in an Asian general population. BMC Nephrol. 2017;18(1):240.
14. Umesawa M, Sairenchi T, Haruyama Y, Nagao M, Yamagishi K, Irie F, et al. Validity of a Risk Prediction Equation for CKD After 10 Years of Follow-up in a Japanese Population: The Ibaraki Prefectural Health Study. Am J Kidney Dis. 2018;71(6):842–50.
15. Dunkler D, Gao P, Lee SF, Heinze G, Clase CM, Tobe S, et al. Risk Prediction for Early CKD in Type 2 Diabetes. Clin J Am Soc Nephrol. 2015;10(8):1371–9.
16. Low S, Lim SC, Zhang X, Zhou S, Yeoh LY, Liu YL, et al. Development and validation of a predictive model for Chronic Kidney Disease progression in Type 2 Diabetes Mellitus based on a 13-year study in Singapore. Diabetes Res Clin Pract. 2017;123:49–54.
17. Riphagen IJ, Kleefstra N, Drion I, Alkhalaf A, van Diepen M, Cao Q, et al. Comparison of Methods for Renal Risk Prediction in Patients with Type 2 Diabetes (ZODIAC-36). PLoS ONE. 2015;10(3):e0120477.
18. Hayes AJ, Leal J, Gray AM, Holman RR, Clarke PM. UKPDS Outcomes Model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia. 2013;56(9):1925–33.
19. Jardine MJ, Hata J, Woodward M, Perkovic V, Ninomiya T, Arima H, et al. Prediction of Kidney-Related Outcomes in Patients With Type 2 Diabetes. Am J Kidney Dis. 2012;60(5):770–8.
20. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomised trials. Lancet Diabetes Endocrinol. 2017;5(10):788–98.
21. Wan EYF, Fong DYT, Fung CSC, Yu EYT, Chin WY, Chan AKC, et al. Prediction of new onset of end stage renal disease in Chinese patients with type 2 diabetes mellitus – a population-based retrospective cohort study. BMC Nephrol. 2017;18(1):257.
22. Elley CR, Robinson T, Moyes SA, Kenealy T, Collins J, Robinson E, et al. Derivation and Validation of a Renal Risk Score for People With Type 2 Diabetes. Diabetes Care. 2013;36(10):3113–20.
23. UK Biobank - UK Biobank [Internet]. [cited 2021 May 14]. Available from: https://www.ukbiobank.ac.uk/
24. UK Biobank. Data providers and dates of data availability [Internet]. [cited 2024 Dec 23]. Available from: https://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=Data_providers_and_dates
25. Radcliffe Department of Medicine. UKPDS Outcomes Model V2 (OM2.) [Internet]. [cited 2023 Oct 31]. Available from: https://www.rdm.ox.ac.uk/about/our-clinical-facilities-and-mrc-units/DTU/software/outcomes
26. Liu H, Zhang L, Xu F, Li S, Wang Z, Han D, et al. Establishment of a prognostic model for patients with sepsis based on SOFA: a retrospective cohort study. J Int Med Res. 2021;49(9):03000605211044892.
27. Kaka AS, Landsteiner A, Ensrud KE, Logan B, Sowerby C, Ullman K, et al. Risk prediction models for diabetic foot ulcer development or amputation: a review of reviews. J Foot Ankle Res. 2023;16(1):13.
28. 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.
29. Junaid KP, Kiran T, Gupta M, Kishore K, Siwatch S. How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. Popul Health Metrics. 2025;23(1):2.
30. Statistical Analysis with Missing Data -, Roderick JA, Little DB, Rubin. - Google Book [Internet]. [cited 2025 Feb 7]. Available from: https://books.google.com.hk/books?hl=zh-CN&lr=&id=BemMDwAAQBAJ&oi=fnd&pg=PR11&dq=Roderick+J.+A.+Little+DBR.+Statistical+Analysis+with+Missing+Data,+2nd+Edition.&ots=FCBWa2GY_U&sig=4JZcc_ko5CUObzxWYrZ1Ej3Z2bA&redir_esc=y#v=onepage&q&f=false
31. Jr FEH, Hmisc: Harrell Miscellaneous [Internet]. 2023 [cited 2023 Mar 22]. Available from: https://CRAN.R-project.org/package=Hmisc
32. van Buuren S, Groothuis-Oudshoorn K, Vink G, Schouten R, Robitzsch A, Rockenschaub P et al. mice: Multivariate Imputation by Chained Equations [Internet]. 2024 [cited 2025 Feb 7]. Available from: https://cran.r-project.org/web/packages/mice/index.html
33. Paik JM, Patorno E, Zhuo M, Bessette LG, York C, Gautam N, et al. Accuracy of identifying diagnosis of moderate to severe chronic kidney disease in administrative claims data. Pharmacoepidemiol Drug Saf. 2022;31(4):467–75.
34. KDIGO EXECUTIVE COMMITTEE. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int. 2024;105(4):A1.
35. Ko S, Venkatesan S, Nand K, Levidiotis V, Nelson C, Janus E. International statistical classification of diseases and related health problems coding underestimates the incidence and prevalence of acute kidney injury and chronic kidney disease in general medical patients. Intern Med J. 2018;48(3):310–5.
36. James F, Nicholas W, Bisher K, Damian F, Timothy E. The body composition and excretory burden of lean, obese, and severely obese individuals has implications for the assessment of chronic kidney disease. Kidney Int. 2014;86(6):1221–8. https://doi.org/10.1038/ki.2014.112.
37. Bello AK, Okpechi IG, Levin A, Ye F, Damster S, Arruebo S, et al. An update on the global disparities in kidney disease burden and care across world countries and regions. Lancet Global Health. 2024;12(3):e382–95.

## Publisher's note