Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hiding in plain sight: finding undiagnosed chronic kidney disease without kidney function data through a machine learning model

Jose David Amorocho-Morales
Corpus AI
Santiago Pelaez
Corpus AI
Juan Felipe Forero
Corpus AI
Jesús Liscano
Corpus AI
Juan Esteban Correa-Morales
juan.correa@ucl.ac.uk

University College London

Research Article

Keywords: Chronic kidney disease, Risk prediction model, External validation, Machine learning, Population Health

Posted Date: September 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10543145/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. JDA-M, SP, JFF and JEC-M are co-founders and shareholders of CorpusAI, the company that holds the algorithm evaluated in this study, which they originally developed as an academic research project. Authors MR and JL are employed by CorpusAI. To mitigate the potential for bias, outcome labels were generated independently of the model developers: eGFR-based CKD status was derived from creatinine measurements in the All of Us electronic health records, computed through a pipeline separate from the feature-generation code. The model was applied without any access to outcome data, all analyses were pre-specified and frozen before any performance metric was computed, and the full pipeline was executed as deterministic, reproducible code.

Abstract

Background

Most chronic kidney disease (CKD) is undiagnosed, yet guideline-supported, low-cost therapy now makes earlier detection actionable. Risk models could prioritise whom to test, but their performance rarely transports across populations. We externally validated a Latin-America-derived, hemodynamic-variable-free case-finding model for occult CKD in a diverse United States cohort.

Methods

Following a pre-specified, frozen protocol and TRIPOD + AI reporting, we applied the model, a gradient-boosted classifier taking 18 routine predictors and no kidney-function input, without any refitting to adults in the All of Us Research Program with a valid outpatient creatinine and no recorded CKD diagnosis. The reference outcome was occult reduced kidney function (anchored, race-free CKD-EPI 2021 eGFR < 45 mL/min/1.73 m²), with a secondary threshold at < 60 in the identical denominator. We assessed discrimination (AUROC), case concentration, net benefit, and calibration before and after a pre-specified two-parameter logistic recalibration, overall and across subgroups of race/ethnicity, sex, age, and household income.

Results

Among 228,346 eligible participants (63.4% female), 1,234 (0.54%) had occult eGFR < 45. The model discriminated with an AUROC of 0.79 (95% CI 0.78–0.80) and concentrated cases steeply: the top 5%, 10%, and 20% of predicted risk captured 23.3%, 40.2%, and 60.7% of all cases, with roughly fourfold enrichment in the top decile. Net benefit exceeded age-only ranking and screen-all across the threshold range appropriate to an inexpensive confirmatory test. Absolute calibration did not transport (expected-to-observed ratio 2.36) but was fully corrected by recalibration (ratio 1.00) with discrimination unchanged. Calibration error was heterogeneous across race/ethnicity and income, tracking differences in outcome prevalence.

Conclusions

A frozen case-finding model transported its risk ordering, though not its probability scale, to a diverse US population, identifying undiagnosed moderate-to-severe CKD among individuals whose kidney function had not been acted upon. Stratified recalibration, not global adjustment, is required for equitable deployment.

Introduction

Chronic kidney disease (CKD) affects an estimated 850 million people worldwide and is among the fastest-growing contributors to the global disease burden (1). Global projections place its prevalence at 436.6 million cases across 31 countries by 2027 (2). Its most consequential epidemiological feature is not its frequency but its invisibility: in the same projection, close to 80% of cases remain undiagnosed (2). The multinational REVEAL-CKD study, designed specifically to quantify undiagnosed stage 3 CKD, confirmed that this diagnostic gap is systematic and consistent across health systems (3). Early detection is a stated priority of current practice, reflected in the 2024 KDIGO guideline's emphasis on timely identification and risk-based evaluation (4).

This gap carries measurable and largely avoidable clinical consequences. In a population-wide analysis of the Stockholm and Skåne regions, only one third of patients with laboratory-confirmed CKD received a diagnosis within the following five years; under simulations of optimal treatment, between one fifth and two fifths of deaths, cardiovascular deaths, and major adverse cardiovascular events would have been preventable had guideline-directed therapy been started early (5). Undiagnosed CKD is therefore a problem of case-finding rather than of therapeutic availability: the window for intervention exists but remains closed because the disease is never identified.

Universal population-wide screening, however, is not cost-effective. A systematic review of economic evaluations found that CKD screening is cost-effective in high-risk groups but ceases to be so in populations without diabetes or hypertension, where incremental cost-effectiveness ratios rise above US$200,000 per quality-adjusted life-year (6). The efficiency of screening thus depends on the ability to target it: to distinguish, within an undiagnosed population, whom to prioritize. Risk-factor-based screening identifies a substantial share of hidden cases, yet considerable heterogeneity persists across studies in definitions, thresholds, and performance (7).

The obstacle to deploying such a model at scale is transportability. Clinical prediction models rarely retain their performance across populations with different demographic structure or baseline risk, and those derived in high-income countries often fail to transport to lower-income settings (8). Discrimination and calibration, however, do not degrade equally. Absolute calibration depends on local incidence and typically requires recalibration, whereas the risk ordering that underpins screening prioritization tends to be preserved. Whether a model derived in Latin America retains this prioritization capacity in a large, diverse United States population has scarcely been examined. We therefore externally validate, in the All of Us Research Program (9), a case-finding model for undiagnosed CKD originally developed in a Latin American cohort. We hypothesized, a priori, that the model would retain useful prioritization capacity to direct screening in both a high-risk and a general-population cohort, even where absolute calibration requires local adjustment.

Methods

Study design and reporting

We conducted an external validation of a previously developed clinical prediction model for undiagnosed chronic kidney disease (CKD), following the TRIPOD + AI reporting guideline for prediction-model studies using regression or machine learning (10, 11). The study was a retrospective, cross-sectional validation. The model was applied to an independent target population without any refitting, feature re-selection, or retraining. The only model-directed intervention performed after scoring was a pre-specified post-hoc recalibration, described below and reported separately from the frozen model (12). The analysis protocol was pre-specified and frozen before any performance metric was computed. To support reproducibility and independent scrutiny, the full analytic pipeline is openly available for review.

Data source and setting

Validation was performed in the All of Us Research Program, a national United States cohort of more than one million participants, over 80% of whom belong to communities historically underrepresented in biomedical research, with linked electronic health records (EHR) (9). We used the Registered Tier Curated Data Repository (CDR v8, release R2024Q3R9), in which EHR data are harmonized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model v5.3. The Registered Tier provides the full EHR domains required for this analysis.

Model under validation

The index model is a case-finding model for occult CKD developed in a Latin American cohort. It estimates the probability that an individual without a recorded CKD diagnosis has moderately-to-severely reduced kidney function (estimated glomerular filtration rate eGFR < 45 mL/min/1.73 m²). It is a gradient-boosted tree classifier (XGBoost) wrapped in an isotonic probability calibrator, takes 18 input features, and, by design, uses no kidney-function variable at inference, so that its predictions are informative for case-finding among individuals in whom kidney function has not been measured or acted upon.

Study population and cohorts

Eligible participants were adults aged 18–89 years with at least one valid outpatient serum creatinine measurement on or before the census date (31 December 2024) and no prior recorded diagnosis of CKD. Prevalent CKD was excluded by removing any participant with a CKD diagnosis code recorded on or before the census date.

Two cohorts were pre-specified, mirroring the two deployment scenarios motivated in the Introduction: (i) the general undiagnosed cohort (all eligible participants), and (ii) a high-risk stratum defined a priori by prevalent diabetes and/or hypertension — the groups in which population screening has been shown to be cost-effective. All primary analyses were performed separately in each cohort.

Outcome

The reference label was computed independently of the feature pipeline. For each participant we assembled the dated series of valid creatinine values up to the census date and computed eGFR with the 2021 CKD-EPI creatinine equation without a race coefficient (Inker et al., 2021; Delgado et al., 2022), applying the sex-specific coefficients ($\kappa = 0.7$, $\alpha = -0.241$, and a multiplier of 1.012 in women; $\kappa = 0.9$, $\alpha = -0.302$ in men)(13, 14).

To prevent misclassification from acute kidney injury (AKI), we applied a pre-specified acute-change rule flagging measurements consistent with AKI (absolute short-interval creatinine rise, a ≥ 1.5× ratio increase within a defined interval, or a transient eGFR drop with subsequent recovery), and anchored each participant's kidney function to a stable eGFR estimate, corroborated by a concordant stable value where available.

Participants whose only stable estimate was an isolated ("singleton") value in a pre-specified borderline band, without albuminuria, were placed in an indeterminate quarantine and excluded from labelling, to avoid classifying ambiguous cases. The quarantine band was fixed a priori at 45–59 mL/min/1.73 m² and, importantly, was parameterized independently of the outcome threshold, so that the analytic denominator is identical across the primary and sensitivity outcome definitions and the two are directly comparable.

The primary binary outcome was y = 1 if the anchored eGFR < 45 mL/min/1.73 m², and 0 otherwise — the threshold for which the model was developed. A pre-specified secondary outcome definition at eGFR < 60 mL/min/1.73 m² (CKD stage G3a or worse) was evaluated in the same denominator to test the sensitivity of conclusions to the severity threshold.

Predictors

The 18 predictors were reconstructed in OMOP from EHR and program survey data, in the exact order expected by the artifact: age, sex, systolic and diastolic blood pressure, body-mass index, diabetes and hypertension indicators, years since diabetes and hypertension onset, fasting glucose, glycated haemoglobin, baseline LDL and HDL cholesterol, baseline weight, ordinal smoking and alcohol status, pulse pressure, and number of antihypertensive agents. Consistent with the model's deployment specification, missing predictor values were passed to the model as zero, the same convention used at training. The impact of this convention was examined in a complete-case sensitivity analysis.

Statistical analysis

The primary metric was the area under the receiver-operating-characteristic curve (AUROC), with 95% confidence intervals from 1,000 bootstrap resamples. Because the outcome is prevalent and binary, a standard (non–time-dependent) AUROC was used. AUC-PR is reported alongside. Because the intended use

is to prioritize whom to screen, we quantified the concentration of true cases within the highest-risk strata: the proportion of all cases captured within the top 5%, 10%, and 20% of predicted risk, and observed prevalence across risk deciles. To assess whether the model's utility survives the trade-off between detected cases and unnecessary tests, we computed net benefit across probability thresholds from 0.001 to 0.10 for four strategies: the model, an age-only ranking, screening everyone, and screening no-one, and identified the dominant strategy at each threshold (15). The threshold range was chosen to span plausible willingness-to-test for an inexpensive confirmatory measurement (a serum creatinine).

We report the calibration-in-the-large intercept, the calibration slope and its accompanying joint intercept, the ratio of expected to observed cases (E/O), and the Brier score, together with a decile-based calibration curve. The intercept-in-the-large and the joint intercept are distinct quantities and are labelled separately throughout.

Anticipating a priori that absolute calibration would require local adjustment while risk ordering would be preserved, we pre-specified a two-parameter logistic recalibration (intercept + slope) of the model's log-odds, applied both overall and stratified by subgroup, and report performance before and after (16). Discrimination is invariant to this monotone transformation and is reported unchanged.

To assess transportability across the cohort's diversity, discrimination (AUROC) and calibration (E/O) were computed within subgroups of race/ethnicity, sex, age band, and a socioeconomic axis derived from the program survey item on annual household income. The socioeconomic axis was pre-specified to be reported only if coverage in the analytic cohort reached 30%; coverage and the resulting decision are reported explicitly. Subgroups with fewer than 20 participants or fewer than 20 events were suppressed per the disclosure rule; between-subgroup gaps in AUROC and the ratio of maximum to minimum E/O were summarized, with a subgroup considered to fail if discrimination collapsed toward chance.

Missing data

Predictor missingness was handled with the model's native training-time convention (zero substitution after numeric coercion) for the primary analysis; a complete-case sensitivity analysis characterized its influence. Outcome labelling required at least one valid stable creatinine-derived eGFR; participants without an anchorable estimate were not labelled.

Reproducibility, governance, and software

All extraction and analysis were executed inside the secure All of Us Researcher Workbench. The frozen analysis protocol, per-notebook provenance manifests, and an append-only run log recorded every artifact produced. Only aggregate outputs with cell counts ≥ 20 were exported; no participant-level features, scores, labels, or cohort files left the environment. Analyses used Python with scikit-learn 1.7.2 and XGBoost 3.0.3. The study was conducted under All of Us Registered Tier access governed by the program's Data Use and Registration Agreement.

Results

Cohort and prevalence of occult CKD

Among adults aged 18–89 years in the All of Us Registered Tier with at least one valid outpatient creatinine measurement and no prior recorded CKD diagnosis, 228,346 participants were eligible for the general undiagnosed cohort after applying the acute-kidney-injury rule and the indeterminate quarantine (45–59 mL/min/1.73 m²). The cohort was 63.4% female. Using the anchored, race-free CKD-EPI 2021 eGFR with sex-specific coefficients, 1,234 participants (0.54%) had occult moderate-to-severe reduction in kidney function (eGFR < 45 mL/min/1.73 m²). This prevalence falls within the 0.45–0.90% range reported for undiagnosed stage G3b+ disease in screening populations (17–19). Applying the pre-specified secondary threshold in the same denominator, 9,002 participants (3.94%) had eGFR < 60 mL/min/1.73 m².

Discrimination and precision

At the primary threshold, the model discriminated occult CKD in the general population with an AUROC of 0.79 (95% CI 0.78–0.80)(Fig. 1A). At the pre-specified secondary threshold of eGFR < 60, the model retained good discrimination in absolute terms (AUROC 0.76, 95% CI 0.75–0.76). In the high-risk stratum defined a priori by prevalent diabetes and/or hypertension (n = 112,721; 888 cases; prevalence 0.79%), the model achieved an AUROC of 0.74 (95% CI 0.72–0.75), below the general-cohort value, as expected in a more homogeneous population. The model concentrated true cases steeply within its highest-risk strata. The top 5%, 10%, and 20% of predicted risk captured 23.3%, 40.2%, and 60.7% of all occult-CKD cases, respectively (Fig. 1B). The top decile alone held four times its proportional share of cases (~ 4× enrichment over random selection), and screening one in five prioritized individuals captured three fifths of all cases. This is the operationally relevant result. In the population for which guideline bodies already consider screening cost-effective, and therefore the population in which a case-finding tool would actually be deployed, age carries less of the discriminative burden and the model contributes proportionally more. See Tables 1 and 2.

Table 1
Model performance by cohort and outcome threshold.

Cohort	Outcome threshold	N	Events	Prevalence, %	AUROC (95% CI)	AUC-PR	Capture top 5%, %	Capture top 10%, %	Capture top 20%, %	Joint intercept	Calibration slope	E/O
General undiagnosed	eGFR < 45	228,346	1,234	0.54	0.79 (0.78–0.80)	0.019	23.34	40.19	60.70	-1.578	0.812	2.356
General undiagnosed	eGFR < 60	228,346	9,002	3.94	0.76 (0.75–0.76)	0.107	18.46	32.96	53.82	0.148	0.725	0.323
High-risk (DM and/or HTN)	eGFR < 45	112,721	888	0.79	0.74 (0.72–0.75)	0.020	18.13	31.08	50.45	-1.936	0.687	2.235
High-risk (DM and/or HTN)	eGFR < 60	112,721	6,695	5.94	0.71 (0.70–0.71)	0.1214	14.19	25.32	44.47	-0.185	0.592	0.296

Table 2
Global two-parameter recalibration. Discrimination is invariant to this monotone transformation and is shown before and after to document that invariance. The complete-case column reports the pre-specified sensitivity analysis on the predictor-missingness convention.

Outcome threshold	Complete cases, n	Joint intercept before	Slope before	E/O before	Brier before	Joint intercept after	Slope after	E/O after	Brier after	AUROC before	AUROC after
eGFR < 45	87,096	-1.578	0.812	2.356	0.0056	-0.167	0.966	1.000	0.0053	0.7896	0.7896
eGFR < 60	87,096	0.148	0.725	0.323	0.0376	0.000	1.000	1.000	0.0365	0.7560	0.7560

Decision-curve analysis

Across probability thresholds of 0.001, 0.002, 0.005, and 0.01, the model achieved the highest net benefit of the four strategies for the primary outcome, exceeding both an age-only ranking and screening everyone (see Fig. 2 Decision-curve analysis of primary outcome). At a threshold of 0.005, for example, net benefit was 0.00227 for the model versus 0.00041 for both age-only and screen-everyone. At thresholds of 0.02 and above, no strategy improved on screening no-one, which is the arithmetically expected consequence of a 0.54% prevalence: a clinician unwilling to test 50 people to find one case cannot benefit from any prioritization rule. The model's advantage therefore lies squarely in the range appropriate to an inexpensive confirmatory test, which is the setting for which it was designed. See Tables 1 and 2.

Calibration and recalibration

Consistent with our a priori hypothesis, absolute calibration did not transfer: the model systematically over-predicted risk, with an expected-to-observed ratio (E/O) of 2.36, an intercept-in-the-large of − 0.887, a joint intercept of − 1.578, and a calibration slope of 0.812 (Brier 0.0056). A pre-specified two-parameter logistic recalibration corrected this fully, yielding E/O 1.00, joint intercept − 0.167, slope 0.97, and Brier 0.0053, while (as expected for a monotone transformation) discrimination was unchanged (AUROC 0.79). The dissociation is the central quantitative finding: the risk ordering that underpins screening prioritization transported intact, whereas the probability scale required, and was amenable to, a light local adjustment. At the secondary threshold the direction reverses, as it must — the model under-predicted a sevenfold more frequent outcome (E/O 0.32, slope 0.725) — and recalibration again restored E/O to 1.00 without altering AUROC.

Subgroup performance and fairness

Discrimination was broadly stable across subgroups meeting the disclosure-driven minimum, with maximum between-subgroup AUROC gaps of 0.094 across race/ethnicity (5 evaluable subgroups), 0.044 across sex (2), 0.13 across age bands (4), and 0.110 across the socioeconomic axis (11). No subgroup collapsed toward chance. The socioeconomic axis was evaluable: the household-income survey item covered 87.1% of the analytic cohort, far exceeding the 30% pre-specified minimum, yielding 11 income categories. See Table 3.

Table 3
Cases captured by the model versus an age-only ranking, at matched screening volumes.

Outcome threshold	Screening fraction, %	Individuals flagged, n	Cases captured by model, %	Cases captured age-only ranking, %	Absolute difference, pp	Ratio model / age	Additional cases per 1,000 screened
eGFR < 45	5	11,576	23.34	23.18	+ 0.16	1.007	+ 0.17
eGFR < 45	10	22,842	40.19	33.55	+ 6.64	1.198	+ 3.59
eGFR < 45	20	45,898	60.70	51.86	+ 8.84	1.170	+ 2.38
eGFR < 60	5	11,576	18.46	21.91	-3.45	0.843	-26.83
eGFR < 60	10	22,842	32.96	34.33	-1.37	0.960	-5.40
eGFR < 60	20	45,898	53.82	54.70	-0.88	0.984	-1.73

Calibration, in contrast, varied systematically and in a direction opposite to the usual expectation. The Black or African American subgroup was the only subgroup already well-calibrated (E/O 1.00; n = 32,972; 350 events), whereas in White participants the model over-predicted by a factor of 3.14 (n = 128,850; 625 events), and in Hispanic/Latino participants by 2.68. The explanation is prevalence: occult CKD was more than twice as frequent in the Black or African American subgroup (1.06%) as in the White subgroup (0.49%), so the absolute risk level learned in the Latin American development population happened to match the former. The ratio of maximum to minimum E/O across race/ethnicity was 3.14, and across the socioeconomic axis 2.86 (E/O 1.40 in the lowest income band versus 3.99 in the highest). For more details among comparison between sex, age band, ethnicity, and income see Supplementary Table 1.

Sensitivity analyses

Restricting to participants with complete predictor data (n = 87,096; 38.1% of the cohort) reproduced the primary findings, indicating that the training-time zero-imputation convention did not drive the result.

Discussion

In a large, diverse United States cohort, a case-finding model for occult CKD developed in Latin America retained clinically useful discrimination (AUROC 0.790) and, more importantly, useful risk-prioritization capacity (top-decile enrichment ≈ 4×; 20% more cases captured than an age-only rule at a 10% screening volume), with incremental value over strong demographic baselines that was larger, not smaller, in the high-risk stratum where screening is already indicated (ΔAUC + 0.041). Net benefit exceeded all comparator strategies across the threshold range appropriate to an inexpensive confirmatory test. The model's only transportability failure was in absolute calibration (anticipated, and fully corrected by two-parameter recalibration without loss of discrimination) with the important caveat that calibration error was heterogeneous across race/ethnicity and income, so that stratified rather than global recalibration is required.

This pattern, discrimination that transports while absolute calibration does not, is not incidental to this model but is the expected behaviour of any risk score moved across populations, and the methodological literature now treats the two properties as separable deployment targets rather than a single notion of "validity." Models routinely achieve strong internal discrimination yet fail calibration standards when transferred across institutions, a dissociation in which a model that correctly ranks patients by relative risk may simultaneously over- or under-estimate absolute probability by a clinically meaningful margin. This is precisely the setting for which geographic external validation followed by recalibration is recommended, because differences in outcome incidence between development and target populations induce miscalibration even when predictor–outcome relationships are preserved. Our results place a Latin-America-derived case-finding model squarely within this paradigm: the ordering that drives screening prioritisation survived the move to a diverse United States cohort, and the probability scale was restored with a two-parameter adjustment rather than any change to the model itself.

The clinical value of preserved ordering depends entirely on whether earlier ascertainment changes what can be offered to the patient, and here the therapeutic context has shifted decisively in favour of case-finding. Under the 2024 KDIGO guidance, SGLT2 inhibitors are a foundational, strongly recommended therapy to slow CKD progression across a broad range of patients, applying to adults with an eGFR of 20 mL/min/1.73 m² or higher irrespective of diabetes status (4). The EMPA-KIDNEY trial was designed to test kidney-disease progression across a wide range of patients with CKD, extending renoprotective benefit beyond the diabetic and heavily albuminuric groups studied earlier (20). A tool that concentrates undiagnosed moderate-to-severe disease into a small, testable fraction of the population therefore feeds directly into an intervention pathway that did not exist a decade ago. The value of ranking is realised only because acting on the top-ranked patients is now guideline-supported and inexpensive.

Set against the scale of the problem, the case for opportunistic, model-guided ascertainment is difficult to dismiss. In the multinational REVEAL-CKD study, the prevalence of undiagnosed stage 3 CKD ranged from 61.6% in the US, to over 90% in France and Japan, and rose with age (3). The same study identified female sex and stage 3a disease as the characteristics most strongly associated with remaining undiagnosed, the very strata in which a laboratory value already exists but is not acted upon (3). Our model operates on exactly this failure mode: it requires no kidney-function input, so it identifies individuals who warrant a first confirmatory creatinine among those the system has not yet flagged, complementing rather than duplicating existing eGFR-based registries.

These arguments carry greatest force where the diagnostic gap is widest and the tools to close it are scarcest. An estimated 78% of people with CKD live in low- and middle-income countries (LMICs), where health systems face the double burden of communicable and non-communicable disease and where kidney disease remains largely absent from the global policy agenda (21). The infrastructure that our approach presupposes cannot be taken for granted in these settings: at the primary-care level, only 12% of low-income countries have access to serum creatinine measurement with eGFR, and none have access to urine albumin-to-creatinine ratio testing (22). The workforce gap is of the same order (LMICs have roughly 0.6 nephrologists per 100,000 patients with CKD, against 25.7 in high-income countries) so that patients characteristically present late, having never been screened, and at least two million people die annually from untreated kidney failure (21, 23). A case-finding model that runs on variables already captured during routine encounters and that requires no kidney-function measurement to generate its prioritisation is, in principle, well suited to precisely these constraints: it can direct a scarce confirmatory creatinine toward those most likely to benefit rather than presupposing universal testing. That the model originated in a Latin American population, rather than being transplanted wholesale from a high-income setting, is not incidental. Screening tools for CKD have overwhelmingly been derived in high-income cohorts, with few developed or validated in LMIC populations (23–25). Our finding that risk ordering transported to a diverse cohort while the probability scale required local recalibration, suggests that models can cross income and geographic boundaries if calibration is treated as a local, revisable quantity rather than a fixed property inherited from the development setting.

Limitations

Several limitations bound these conclusions. First, this was a cross-sectional external validation against a single anchored eGFR-defined reference, without albuminuria; because staging omitted the ACR axis, we necessarily under-ascertained CKD defined by KDIGO's full criteria and cannot speak to albuminuric

normal-eGFR disease. Second, the reference label itself depends on the CKD-EPI 2021 equation. Third, calibration error was heterogeneous across race/ethnicity and income, so a single global recalibration is insufficient; the stratified recalibration this implies was not computable for the socioeconomic axis, which was added after the primary analysis, and is reported only for discrimination and pre-recalibration calibration. Fourth, discrimination degraded most across age bands, the axis on which the model's incremental value over an age-only rule is intrinsically smallest, and the absolute number of events in some strata fell near the disclosure floor, widening uncertainty. Finally, validation in a single national research cohort with its own recruitment and EHR-capture patterns does not establish transportability to routine-care populations elsewhere. The missing-as-zero convention, although not materially influential in complete-case sensitivity analysis, remains a deployment dependency (26).

Conclusion

A case-finding model for occult CKD developed in Latin America retained clinically useful discrimination and, more importantly, useful risk-prioritisation capacity when applied without refitting to a large, diverse United States cohort, concentrating roughly four times its proportional share of undiagnosed moderate-to-severe disease into the top risk decile. Its only transportability failure was in absolute calibration — anticipated, and fully corrected by two-parameter recalibration without loss of discrimination — with the important caveat that calibration error was heterogeneous across race/ethnicity and income, so that stratified rather than global recalibration is required before deployment. Because the model uses no kidney-function variable at inference and is therefore informative precisely where kidney function has not yet been measured, it offers a practical, low-cost mechanism to direct a first confirmatory creatinine toward those most likely to harbour undiagnosed disease, at a moment when guideline-supported therapy makes earlier detection actionable.

Declarations

Ethics

The study uses de-identified, retrospective real-world data under the participating institutions' governance framework. Both the predictive-model derivation and the deployment and outcome-evaluation components received institutional review board approval by Salud Total ethics committee in June 2025 and All of Us in May 2026, with a waiver of informed consent owing to the use of de-identified, retrospective data.

Declaration of competing interest:

JDA-M, SP, JFF and JEC-M are co-founders and shareholders of CorpusAI, the company that holds the algorithm evaluated in this study, which they originally developed as an academic research project. Authors MR and JL are employed by CorpusAI. To mitigate the potential for bias, outcome labels were generated independently of the model developers: eGFR-based CKD status was derived from creatinine measurements in the All of Us electronic health records, computed through a pipeline separate from the feature-generation code. The model was applied without any access to outcome data, all analyses were pre-specified and frozen before any performance metric was computed, and the full pipeline was executed as deterministic, reproducible code.

Generative AI statement

The authors declare that Claude (Anthropic) was used solely to assist with language editing of the manuscript text. All content generated or edited with this tool was subsequently reviewed, verified, and approved by the authors, who take full responsibility for the accuracy, integrity, and originality of the manuscript, including all data, citations, and references.

Competing Interests

JDA-M, SP, JFF and JEC-M are co-founders and shareholders of CorpusAI, the company that holds the algorithm evaluated in this study, which they originally developed as an academic research project. Authors MR and JL are employed by CorpusAI. To mitigate the potential for bias, outcome labels were generated independently of the model developers: eGFR-based CKD status was derived from creatinine measurements in the All of Us electronic health records, computed through a pipeline separate from the feature-generation code. The model was applied without any access to outcome data, all analyses were pre-specified and frozen before any performance metric was computed, and the full pipeline was executed as deterministic, reproducible code.

Funding:

This study received no external funding.

Author Contribution

JEC-M: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review and editing, Supervision, Project administration. JDA-M: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review and editing. JL: Investigation, Data curation, Writing – review and editing. JFF, SP: Data curation, Writing – review and editing. All authors contributed to the article and approved the submitted version.

Data Availability

The data analysed in this study were obtained from the All of Us Research Program and are not publicly distributable by the authors. All of Us data are available to authorised researchers through the All of Us Researcher Workbench (https://www.researchallofus.org), subject to registration, institutional agreement, and completion of the program's Registered Tier Data Use and Registration Agreement; participant-level data cannot be exported from the secure environment under this agreement. The analyses in this study used the Registered Tier Curated Data Repository (CDR v8, release R2024Q3R9). The full analytic code, frozen protocol, and per-notebook provenance manifests supporting the findings are available from the corresponding author on reasonable request. The index model derived in the Latin American cohort is proprietary to CorpusAI, requests regarding the model should be directed to the corresponding author.

References

1. Jager KJ, et al. A single number for advocacy and communication-worldwide more than 850 million individuals have kidney diseases. Nephrol Dial Transpl. 2019;34:1803–5.
2. Chertow GM, Correa-Rotter R, Eckardt KU, et al. Projecting the clinical burden of chronic kidney disease at the patient level (Inside CKD): a microsimulation modelling study. EClinicalMedicine. 2024;72:102614. https://doi.org/10.1016/j.eclinm.2024.102614.
3. Kushner P, Peach E, Wittbrodt E, et al. Investigating the global prevalence and consequences of undiagnosed stage 3 chronic kidney disease: methods and rationale for the REVEAL-CKD study. Clin Kidney J. 2021;15(4):738–46. https://doi.org/10.1093/ckj/sfab235.
4. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int. 2024;105(4S):S117–314. https://doi.org/10.1016/j.kint.2023.10.018.
5. Sundström J, Norhammar A, Karayiannides S, et al. Are there lost opportunities in chronic kidney disease? A region-wide cohort study. BMJ Open. 2024;14(4):e074064. https://doi.org/10.1136/bmjopen-2023-074064.
6. Yeo SC, Wang H, Ang YG, Lim CK, Ooi XY. Cost-effectiveness of screening for chronic kidney disease in the general adult population: a systematic review. Clin Kidney J. 2023;17(1):sfad137. https://doi.org/10.1093/ckj/sfad137.
7. Korsa A, Tesfaye W, Sud K, Krass I, Castelino RL. Risk factor-based screening for early detection of chronic kidney disease in primary care settings: a systematic review. Kidney Med. 2025;7(4):100979. https://doi.org/10.1016/j.xkme.2025.100979.
8. Carrillo-Larco RM, Moscoso-Porras M, Taype-Rondan A, et al. Systematic review of diagnostic and prognostic models of chronic kidney disease in low-income and middle-income countries. BMJ Open. 2022;12(3):e058733. 10.1136/bmjopen-2021-058733.
9. Denny JC, Rutter JL, Goldstein DB, et al. The All of Us Research Program. N Engl J Med. 2019;381(7):668–76. https://doi.org/10.1056/NEJMsr1809937.
10. Collins GS, Moons KGM, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;16:e078378. https://doi.org/10.1136/bmj-2023-078378.
11. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49–58. 10.1093/ckj/sfaa188.
12. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023;21(1):70. 10.1186/s12916-023-02779-w.
13. Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med. 2021;385(19):1737–49. https://doi.org/10.1056/NEJMoa2102953.
14. Delgado C, Baweja M, Crews DC, et al. A unifying approach for GFR estimation: recommendations of the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. Am J Kidney Dis. 2022;79(2):268–e2881. https://doi.org/10.1053/j.ajkd.2021.08.003.
15. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565–574. 10.1177/0272989X06295361. PMID: 17099194.
16. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. 10.1136/bmj-2023-074819.
17. Bang H, Vupputuri S, Shoham DA, et al. SCreening for Occult REnal Disease (SCORED): a simple prediction model for chronic kidney disease. Arch Intern Med. 2007;167(4):374–81. 10.1001/archinte.167.4.374.
18. Yoo D, Nguyen VK, Maggiore U, Jolliet O. MERWACS: Development and external validation of a non-invasive machine learning tool for identifying subjects to be screened for CKD. PLOS Digit Health. 2026;9(7):e0001486. 10.1371/journal.pdig.0001486.
19. Galbete A et al. (o el autor correspondiente) Risk Prediction Score for Chronic Kidney Disease in Healthy Adults and Adults With Type 2 Diabetes: Systematic Review. *Prev Chronic Dis.* 2023. — verificada en CDC/PCD.
20. The EMPA-KIDNEY Collaborative Group, Herrington WG, Staplin N, Wanner C, et al. Empagliflozin in Patients with Chronic Kidney Disease. N Engl J Med. 2023;388(2):117–27. 10.1056/NEJMoa2204233.
21. Kalyesubula R, Conroy AL, Calice-Silva V, Kumar V, Onu U, Batte A, Kaze FF, Fabian J, Ulasi I. Screening for Kidney Disease in Low- and Middle-Income Countries. Semin Nephrol. 2022;42(5):151315. 10.1016/j.semnephrol.2023.151315.
22. Crews DC, Bello AK, Saadi G, World Kidney Day Steering Committee. Burden, Access, and Disparities in Kidney Disease. Kidney Int Rep. 2019; 4;4(3):372–379. 10.1016/j.ekir.2019.01.011
23. Francis A, Harhay MN, Ong ACM, American Society of Nephrology; European Renal Association; International Society of Nephrology, et al. Chronic kidney disease and the global public health agenda: an international consensus. Nat Rev Nephrol. 2024;20(7):473–85. 10.1038/s41581-024-00820-6.

24. Ke C, Liang J, Liu M, Liu S, Wang C. Burden of chronic kidney disease and its risk-attributable burden in 137 low- and middle-income countries, 1990–2019: results from the Global Burden of Disease study 2019. BMC Nephrol. 2022;23(1):17. 10.1186/s12882-021-02597-3.
25. Kaze AD, Ilori T, Jaar BG, Echouffo-Tcheugui JB. Burden of chronic kidney disease on the African continent: a systematic review and meta-analysis. BMC Nephrol. 2018;19(1):125.
26. Oliver JD 3rd, Nee R, Marneweck H, Banaag A, Koyama AK, Pavkov ME, Koehlmoos TP. Impact of Race-Free Glomerular Filtration Rate Estimations on CKD Prevalence in the US Military Health System: A Retrospective Cohort Study. Kidney Med. 2024; 21;6(8):100861. 10.1016/j.xkme.2024.100861

Figures

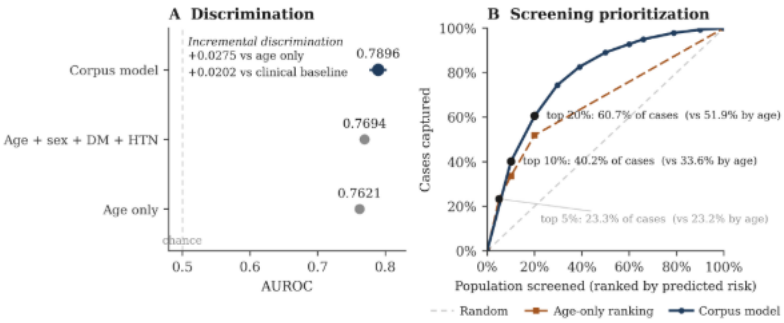


Figure 1

Discrimination and screening prioritization of the occult-CKD model in the All of Us undiagnosed cohort (n = 228,346; 1,234 cases, prevalence 0.54%). (A) Area under the receiver-operating-characteristic curve (AUROC) for the model (0.7896; horizontal bar, 95% bootstrap confidence interval 0.7781–0.8001) compared with two logistic baselines fitted in the target data: age alone (0.7621) and age + sex + diabetes + hypertension (0.7694); the dashed line marks chance (0.5). The model adds discrimination beyond demographic and comorbidity information (ΔAUC +0.0275 and +0.0202). **(B)** Cumulative gains curve: proportion of all true occult-CKD cases captured (y-axis) as an increasing fraction of the population, ranked by predicted risk, is screened (x-axis); the dashed diagonal is random selection and the dotted line is an age-only ranking. Markers indicate the exported operating points — the highest-risk 5%, 10%, and 20% captured 23.3%, 40.2%, and 60.7% of cases, respectively (top-decile enrichment ≈4× over random). Curves and points are reconstructed from decile-level aggregates exported under the disclosure rule (cell counts ≥20).

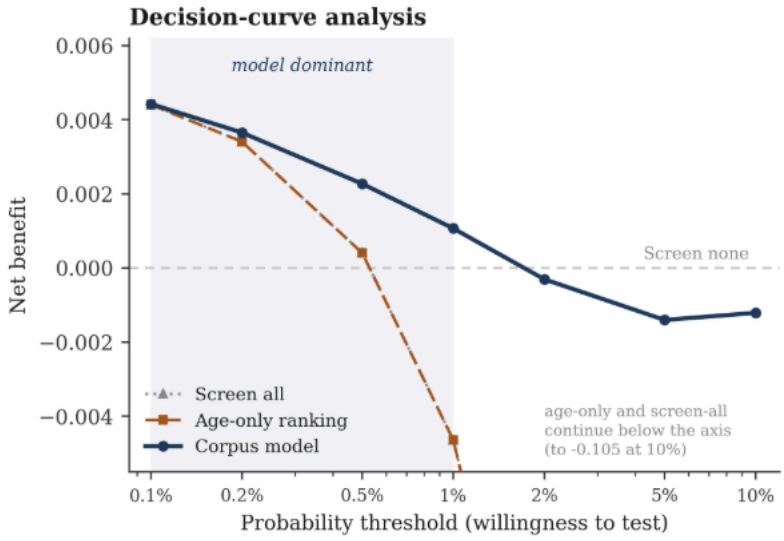


Figure 2

Decision-curve analysis of primary outcome (eGFR < 45 mL/min/1.73 m²). Net benefit (y-axis) across probability thresholds from 0.001 to 0.10 (x-axis, log scale) for four strategies: the model, an age-only ranking, screening all participants, and screening none. The model achieves the highest net benefit at thresholds of 0.001–0.01; at 0.02 and above, screening none dominates, consistent with a prevalence of 0.54%.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial1.docx