

Uncertainty-Aware Multi-Outcome Screening for Undiagnosed Cardiometabolic Disease in the United States Population Using NHANES

Kazi Sabbir Ahmad Nahin

University of Kentucky

Arman Hossen

armanmamun07@gmail.com

Miami University

Abida S Asha

University of Kentucky

K. M. Ashfak Alam Siam

Miami University

Md. Shakil Ahmed

Miami University

Ashiqur Rahman Rony

Government Titumir College

Research Article

Keywords: Cardiometabolic disease, Undiagnosed disease, Screening, Machine learning, Conformal prediction, Uncertainty quantification, Calibration, NHANES, Clinical decision support

Posted Date: August 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10708560/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Cardiometabolic diseases, including diabetes, hypertension, dyslipidemia, and chronic kidney disease, are leading causes of morbidity and mortality, and many affected adults remain undiagnosed until complications appear. Machine learning can support population screening, but model probabilities are often poorly calibrated and give no indication of when a prediction should be trusted. Using the National Health and Nutrition Examination Survey (NHANES) 2011 to 2014, we assembled an analytical cohort of 20,374 adults aged 18 years or older and developed a screening framework that pairs multi-outcome prediction with conformal prediction. Five outcomes representing previously undiagnosed disease were defined by combining the absence of a self-reported physician diagnosis with objective biomarker or examination evidence, and any biomarker used to define an outcome was excluded as a predictor for that outcome. Three nested deployment scenarios represented community, primary-care, and laboratory-supported settings, and six models were compared.

Results

In the analytical cohort, 26.3% of adults had evidence of at least one undiagnosed cardiometabolic condition, and the survey-weighted prevalence of any latent disease was 26.6% (95% CI 25.1 to 28.2). Discrimination improved as clinical information increased; for the composite outcome the Random Forest AUROC rose from 0.789 in the community scenario to 0.913 with examination data and 0.964 with laboratory data. Under routine clinical assessment, tree-ensemble models reached an AUROC of 0.984 for dyslipidemia, 0.977 for hypertension, 0.922 for diabetes, and 0.999 for possible chronic kidney disease risk, and repeated cross-validation confirmed that these estimates were stable across folds. Conformal prediction achieved empirical coverage close to the 90% target for every model, yet the fraction of individuals flagged as uncertain differed widely between models with similar discrimination, ranging from 0.3% for Random Forest to 59.9% for logistic regression on the composite outcome.

Conclusions

Reporting a prediction together with its uncertainty makes it possible to screen for multiple undiagnosed cardiometabolic conditions across realistic care settings while identifying the individuals for whom a screening result is not reliable and confirmatory testing is warranted. The uncertainty rate is a decision-relevant quantity that discrimination metrics alone do not reveal.

Background

Cardiometabolic diseases, including type 2 diabetes, hypertension, dyslipidemia, and chronic kidney disease (CKD), are among the leading contributors to global morbidity, mortality, and health-care expenditure [1–3]. A common feature of these conditions is that they often progress silently. Raised blood glucose, elevated blood pressure, atherogenic lipids, and early kidney injury can persist for years before symptoms or complications appear, so a large share of the adult population meets biochemical or physiological criteria for disease while remaining formally undiagnosed [4–6]. National surveillance has repeatedly shown that a substantial fraction of diabetes remains undiagnosed, that many adults meet contemporary thresholds for elevated blood pressure without being aware of it, and that chronic kidney disease is frequently unrecognized until advanced stages [7–9]. Because even asymptomatic dysglycemia and subclinical vascular injury raise the long-term risk of cardiovascular events, this undiagnosed burden is far from benign [10]. Undiagnosed disease therefore represents a missed opportunity for low-cost intervention, and reducing it is a central objective of preventive medicine and public health screening [11].

Population-based screening is an established strategy for detecting individuals at elevated risk before irreversible damage occurs [11, 12]. Surveys such as the National Health and Nutrition Examination Survey (NHANES) are well suited to studying undiagnosed disease because they combine self-reported diagnostic history with standardized physical examinations and laboratory biomarkers in a nationally representative sample [13]. This design makes it possible to define undiagnosed disease directly, as the co-occurrence of objective disease evidence and the absence of a prior diagnosis, rather than relying on self-report alone.

Machine learning has enabled predictive models that capture nonlinear relationships and interactions among clinical, behavioral, and demographic variables, and such models have improved discrimination for cardiometabolic risk relative to traditional scores [14–17]. Reviews of artificial intelligence in medicine and of machine learning for diabetes describe rapid growth in this area and a broad menu of

algorithms applied to screening and risk stratification [18, 19]. Discrimination alone, however, is not sufficient for safe screening, and the advantage of complex models over a well-specified logistic regression is often smaller than expected once studies are appraised carefully [20]. The probability outputs of flexible models are also frequently miscalibrated, so a reported risk may not correspond to an actual event rate [21, 22]. A point prediction also conveys no information about its own reliability. Two individuals may receive the same predicted probability while one lies in a well-characterized region of the data and the other in a sparse and ambiguous one. In a screening setting, treating those two predictions identically is the failure mode that erodes clinical trust.

Prediction uncertainty has therefore become a central concern in trustworthy artificial intelligence for health care [23, 24]. Conformal prediction offers a model-agnostic solution. Using a held-out calibration set, it converts a base model's scores into prediction sets with finite-sample and distribution-free coverage guarantees [25–27]. For a binary screening task, a conformal prediction set may contain a single label, which is a confident low-risk or high-risk decision, or both labels, which marks the case as uncertain and appropriate for confirmatory testing or referral. This changes the role of the model, so that it decides when a confident decision can be made and when a decision should be deferred. Although conformal prediction remains uncommon in clinical research, recent reviews have begun to catalogue its applications and to argue for wider adoption in medicine [28].

Several gaps remain despite this progress. Most clinical prediction studies focus on a single disease, commonly diabetes, rather than the correlated cluster of cardiometabolic conditions that co-occur in practice, even when trained on the same national surveys [29]. Few studies evaluate how predictive performance and uncertainty behave across realistic deployment tiers, from survey-only community screening to laboratory-supported clinical assessment, and few translate uncertainty estimates into an operational referral signal that could be used at the point of care. The most closely related work developed survey-design-aware neural network risk scores with uncertainty quantification for diabetes in NHANES 2011 to 2014 [30]. The present study extends that direction to a multi-outcome, deployment-tiered, conformal-prediction framework.

A further concern is equity. Widely deployed clinical algorithms can encode and amplify structural bias, as shown by an influential audit of a population health algorithm that systematically underestimated the care needs of Black patients [31]. For a screening tool that decides who is referred for confirmatory testing, uncertainty that is unevenly distributed across demographic groups could widen rather than narrow existing disparities. An uncertainty-aware framework should therefore be evaluated for average coverage and for how its confident and uncertain decisions are distributed across the population.

To address these gaps, we developed an uncertainty-aware screening framework for undiagnosed cardiometabolic disease using NHANES 2011 to 2014. We built a unified framework for the simultaneous screening of multiple undiagnosed cardiometabolic conditions, including a shared-encoder multi-task neural network, rather than a single-disease model. We quantified the incremental value of clinical information by evaluating three deployment scenarios that mirror community, primary-care, and clinical settings, using an explicit rule that excludes any defining biomarker from the predictor set for the corresponding outcome. Finally, we integrated split conformal prediction to translate model outputs into uncertainty-aware prediction sets, and we show that models with near-identical discrimination can differ substantially in how much confident decision-making they can support, which is information that standard accuracy metrics do not reveal but that is decisive for real-world screening.

Methods

Data source and study design

Data were obtained from NHANES, a nationally representative cross-sectional survey conducted by the National Center for Health Statistics (NCHS) of the United States Centers for Disease Control and Prevention (CDC). NHANES uses a complex multistage probability sampling design and collects demographic, socioeconomic, questionnaire, physical examination, and laboratory information from the civilian, non-institutionalized United States population [13]. The development dataset comprised the 2011 to 2012 and 2013 to 2014 cycles. Publicly available demographic, questionnaire, examination, and laboratory files were downloaded from the CDC website and merged on the unique participant identifier (SEQN), so that each participant contributed a single integrated record. Because NHANES uses a complex survey design, the sampling weights, primary sampling units, and strata were retained throughout data preparation.

Study population

Adult participants aged 18 years or older were eligible. Individuals with insufficient information to define the study outcomes were excluded during preprocessing. From 28,879 initial participants, the final analytical cohort comprised 20,374 adults (Fig. 2).

Outcome definitions

The objective was to identify individuals with previously undiagnosed disease, defined as objective disease evidence in the absence of a self-reported physician diagnosis. Four disease-specific outcomes and one composite outcome were constructed, with full thresholds provided in Supplementary Table S1.

Undiagnosed diabetes was defined as no prior physician diagnosis of diabetes together with biomarker evidence, namely hemoglobin A1c of 6.5% or higher or fasting plasma glucose of 126 mg/dL or higher. Undiagnosed hypertension was defined as no prior diagnosis of high blood pressure together with a measured mean systolic blood pressure of 130 mmHg or higher or a mean diastolic blood pressure of 80 mmHg or higher. Undiagnosed dyslipidemia was defined as no prior diagnosis of high cholesterol together with total cholesterol of 240 mg/dL or higher, HDL cholesterol below 40 mg/dL, or triglycerides of 200 mg/dL or higher. Possible CKD risk was defined by a urinary albumin-to-creatinine ratio of 30 mg/g or higher, reflecting albuminuria; it is described as possible CKD risk because definitive staging requires the estimated glomerular filtration rate and confirmation of chronicity. The composite outcome, any latent cardiometabolic disease, was positive if any of the four conditions was present.

To avoid diagnostic leakage, a biomarker used to define an outcome was never used as a predictor for that same outcome. Hemoglobin A1c and fasting glucose were excluded from the diabetes models, measured blood pressure from the hypertension models, total cholesterol, HDL, and triglycerides from the dyslipidemia models, and the albumin-to-creatinine ratio from the CKD-risk models. This ensures that the models perform genuine screening from upstream information rather than recovering the defining test.

Deployment scenarios

Because the information available to a clinician differs by care setting, three nested deployment scenarios were defined, with the full variable list in Supplementary Table S2. Scenario 1 represented community or survey-only screening and used demographic, socioeconomic, and questionnaire variables that can be obtained without examination or laboratory testing, such as age, sex, race and ethnicity, education, income-to-poverty ratio, and medical history. Scenario 2 represented primary-care screening and added routine physical examination measurements, including body mass index, waist circumference, and blood pressure. Scenario 3 represented routine clinical assessment and added commonly available laboratory biomarkers, such as total and HDL cholesterol, triglycerides, and serum creatinine, excluding any biomarker used to define the outcome under study. This nested design allows a direct assessment of the incremental value of increasingly rich and increasingly costly clinical information.

Data integration and preprocessing

Component files were merged on SEQN, and duplicate variables across survey components were removed before merging. Continuous variables were screened for implausible values. Missingness was substantial for several laboratory measures because of the NHANES subsampling design; for example, fasting glucose and triglycerides were measured only in the fasting subsample, as summarized in Supplementary Table S3. For the primary analysis, continuous predictors were imputed using the median and categorical predictors were encoded appropriately to produce a complete, model-ready dataset. Multiple imputation is identified as a planned sensitivity analysis.

Predictive models

Six models spanning linear, tree-ensemble, and neural network families were trained under an identical training and evaluation pipeline for a fair comparison. Logistic regression served as the interpretable linear baseline. Random Forest was a bagged decision-tree ensemble, and XGBoost and LightGBM were gradient-boosted decision-tree ensembles that capture nonlinearities and interactions. The multi-task neural network was a feed-forward network with a shared encoder and separate output heads for each outcome, which allows joint learning across correlated conditions. The residual multi-task neural network extended this architecture with residual connections and a 128-dimensional shared embedding layer, from which learned patient representations were extracted. The neural networks were trained under Scenario 3 to predict all outcomes jointly. Data were partitioned into training and held-out test sets in an approximately 80 to 20 ratio, giving a test sample of 4,075 participants that was held fixed across all models and scenarios. Model hyperparameters were not tuned by an automated search; instead, each model used a fixed, pre-specified configuration selected from established defaults with light manual adjustment on the training data, and the complete configuration of every model, including the learning rate, batch size, number of training epochs, and early-stopping rule for the neural networks, is reported in full in Supplementary Table S4.

Cross-validation for stability

Because the primary evaluation relied on a single held-out split, we additionally performed repeated stratified k-fold cross-validation to verify that discrimination was stable across different partitions of the population rather than an artifact of one split. Under Scenario 3, every model was refitted using five-fold cross-validation repeated three times, giving fifteen held-out evaluations per model and outcome, with the same preprocessing pipelines and hyperparameters used in the main analysis. The mean and standard deviation of AUROC and AUPRC across folds are reported in Supplementary Table S7.

Uncertainty quantification using conformal prediction

Uncertainty was quantified using split conformal prediction applied to each model's predicted probabilities for each outcome under Scenario 3 [25–27]. For each model and outcome, the held-out predictions were split equally into a calibration set and an evaluation set, stratified by outcome. On the calibration set, nonconformity scores were computed as the probability mass assigned to the incorrect label, following adaptive-coverage conformal classification [32]. The conformal threshold was taken as the empirical quantile of the calibration scores at a miscoverage level of 0.10, which corresponds to a target coverage of 90%.

For a new individual with predicted probability p , the prediction set includes the negative label when p is at or below the threshold and the positive label when $1 - p$ is at or below the threshold. This yields four possible prediction sets: a confident low-risk set, a confident high-risk set that indicates referral for confirmatory testing, an uncertain set containing both labels, and, for very rare outcomes, an empty set. We reported, for each model and outcome, the empirical coverage, the average prediction set size, the uncertainty rate defined as the proportion of sets containing both labels, the singleton rate, and the empty rate. The uncertainty rate is the operationally important quantity, because it is the fraction of the population that the screening system would flag for confirmatory testing rather than classify outright.

Evaluation metrics

Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC), the latter being informative under class imbalance. Classification performance was summarized using accuracy, sensitivity, specificity, precision, and F1-score computed at the default probability threshold of 0.50. Because a single fixed threshold is not optimal for every outcome, particularly under the strong class imbalance of the rarer conditions, we additionally report these threshold-dependent metrics at the Youden-optimal and F1-optimal thresholds for every model and outcome in Supplementary Table S6; threshold-independent discrimination and the conformal analysis described below were the primary basis for model comparison. Calibration was assessed using the Brier score and calibration curves. Confidence intervals of 95% for AUROC and AUPRC were estimated by bootstrap resampling. Metrics were computed separately for every model, deployment scenario, and outcome.

Population estimates

Because NHANES uses a complex probability design, population prevalence was estimated using the combined four-year mobile examination center (MEC) weights, obtained by dividing each two-year weight by two. Design-based 95% confidence intervals were computed by Taylor-series linearization that accounts for stratification and clustering.

Clinical utility and learned representations

Decision-curve analysis compared the net benefit of each model against treat-all and treat-none strategies across a range of threshold probabilities [33]. To examine what the neural networks learned, the 128-dimensional embeddings from the residual multi-task network were projected to two dimensions using Uniform Manifold Approximation and Projection (UMAP) and inspected for structure relating to the number of co-occurring latent conditions [34]. To quantify this structure rather than rely on visual impression, we computed the silhouette coefficient of the full 128-dimensional embedding, using the number of co-occurring undiagnosed conditions as the grouping label; the analysis is summarized in Supplementary Figure S1.

Statistical software

All analyses were implemented in Python. Outcome construction, feature engineering, model training, conformal calibration, and figure and table generation were organized as scripted, reproducible modules.

Ethical considerations

NHANES data are publicly available and de-identified before release. The NHANES protocol was approved by the NCHS Research Ethics Review Board, and participants provided written informed consent at data collection. The present analysis used publicly accessible secondary data and required no additional recruitment or consent.

Results

Study population and workflow

The overall study workflow, covering dataset integration, outcome derivation, deployment-scenario construction, modeling, and uncertainty quantification, is summarized in Fig. 1. Participant selection is shown in Fig. 2. Of 28,879 NHANES 2011 to 2014 participants, 20,374 adults aged 18 years or older with sufficient information for outcome definition formed the analytical cohort.

Baseline characteristics are summarized in Table 1. The cohort had a mean age of 53.6 years (SD 18.6), with 33.0% aged 65 years or older and a near-balanced sex distribution of 47.2% male and 52.8% female. The mean body mass index was 29.6 kg/m² (SD 7.1) and the mean waist circumference was 101.1 cm (SD 16.0), indicating a population with a high burden of adiposity. Mean systolic and diastolic blood pressures were 125.5 mmHg (SD 18.5) and 69.2 mmHg (SD 13.0), and the median income-to-poverty ratio was 1.79 (IQR 1.02 to 3.48).

Table 1
Baseline characteristics of the analytical cohort (N = 20,374).

Characteristic	Overall cohort (N = 20,374)
Age, years, mean (SD)	53.6 (18.6)
Age ≥ 65 years, n (%)	6,718 (33.0)
Male, n (%)	9,624 (47.2)
Female, n (%)	10,750 (52.8)
BMI, kg/m ² , mean (SD)	29.6 (7.1)
Waist circumference, cm, mean (SD)	101.1 (16.0)
Systolic blood pressure, mmHg, mean (SD)	125.5 (18.5)
Diastolic blood pressure, mmHg, mean (SD)	69.2 (13.0)
Poverty-income ratio, median (IQR)	1.79 (1.02–3.48)

Prevalence of undiagnosed cardiometabolic conditions

The prevalence of each undiagnosed outcome in the analytical cohort is shown in Fig. 3 and Table 2. The composite outcome, any latent cardiometabolic disease, was the most common, affecting 26.3% of participants (5,360 of 20,374). Among the individual conditions, undiagnosed dyslipidemia was the most prevalent at 13.0%, followed by undiagnosed hypertension at 12.6% and undiagnosed diabetes at 4.2%. Possible CKD risk was uncommon at 0.34%.

Survey-weighted estimates using the combined four-year MEC weights were consistent with the unweighted figures (Table 2). The design-weighted prevalence was 26.6% (95% CI 25.1 to 28.2) for any latent cardiometabolic disease, 13.7% (95% CI 12.6 to 14.8) for undiagnosed dyslipidemia, 13.5% (95% CI 12.2 to 14.7) for undiagnosed hypertension, 3.1% (95% CI 2.6 to 3.6) for undiagnosed diabetes, and 0.21% (95% CI 0.10 to 0.32) for possible CKD risk. The finding that roughly one in four United States adults shows objective evidence of a previously unrecognized cardiometabolic abnormality despite reporting no prior diagnosis indicates a large and actionable screening gap.

Table 2

Prevalence of undiagnosed cardiometabolic outcomes in the analytical cohort: unweighted proportions and survey-weighted estimates with 95% confidence intervals.

Outcome	Cases (n)	Unweighted prevalence (%)	Survey-weighted prevalence (%)	95% CI
Undiagnosed diabetes	865	4.25	3.11	2.61–3.61
Undiagnosed hypertension	2564	12.58	13.46	12.23–14.69
Undiagnosed dyslipidemia	2645	12.98	13.72	12.61–14.83
Possible CKD risk	69	0.34	0.21	0.10–0.32
Any latent cardiometabolic disease	5360	26.31	26.61	25.07–28.15

Discrimination improves with clinical information

Across nearly all outcomes, discrimination improved as examination and laboratory variables were added from Scenario 1 to Scenario 3 (Fig. 4, Table 3). For the composite outcome, the Random Forest AUROC increased from 0.789 (95% CI 0.773 to 0.806) under community screening, to 0.913 (95% CI 0.902 to 0.922) with physical examination, and to 0.964 (95% CI 0.959 to 0.970) with routine laboratory information. Comparable gains were seen for the individual outcomes; for example, the Random Forest AUROC for undiagnosed diabetes rose from 0.838 to 0.901 to 0.922, and for undiagnosed hypertension from 0.765 to 0.973 to 0.977 across the three scenarios.

Model comparison under routine clinical assessment

Under Scenario 3, tree-ensemble models delivered the strongest discrimination and consistently outperformed logistic regression (Table 3). Random Forest and LightGBM were the leading models across most outcomes. The AUROC reached 0.984 for undiagnosed dyslipidemia for both models, 0.977 for undiagnosed hypertension for Random Forest and 0.972 for LightGBM, 0.922 for undiagnosed diabetes for Random Forest, and 0.999 for possible CKD risk for both models. For the composite outcome, Random Forest achieved an AUROC of 0.964 and an AUPRC of 0.912, compared with 0.708 and 0.435 for logistic regression. The multi-task neural networks achieved competitive discrimination that remained below the strongest tree ensembles, with composite AUROCs of 0.864 for the multi-task network and 0.921 for the residual multi-task network, while providing a single unified model for all outcomes and, in the residual variant, a shared latent representation of cardiometabolic burden. The lower precision for CKD risk reflects the rarity of that outcome rather than poor ranking, as shown by the near-perfect AUROC.

Because the classification metrics in Table 3 were computed at the default probability threshold of 0.50, some threshold-dependent values reflect that operating point rather than a difference in ranking. For example, Random Forest attained a specificity and precision of 1.000 for undiagnosed diabetes under Scenario 3 because, at a threshold of 0.50, it labeled only 107 of 4,075 test participants as positive and all of these were true positives, at the cost of a sensitivity of 0.618. At a screening-appropriate operating point defined by Youden's index (threshold 0.20), the same model reached a sensitivity of 0.763 and a specificity of 0.933 (Supplementary Table S6). Such threshold effects, which are most pronounced for the rarer outcomes, show why threshold-independent discrimination and the conformal analysis described below, rather than metrics at any single fixed cut-off, form the basis of our model comparison.

Repeated stratified five-fold cross-validation confirmed that these results are stable across different partitions of the cohort rather than a feature of the single held-out split (Supplementary Table S7). For the composite outcome, the cross-validated AUROC was 0.967 (SD 0.002) for Random Forest, 0.955 (SD 0.002) for LightGBM, 0.889 (SD 0.006) for XGBoost, and 0.702 (SD 0.010) for logistic regression, each within 0.01 of the corresponding single-split estimate. Fold-to-fold variability was small across outcomes, with an AUROC standard deviation below 0.02 for every model on all outcomes except the rare possible-CKD-risk outcome, where it was larger for some models (up to 0.034), as expected from its low event count of 69. The multi-task neural networks showed the same pattern of small fold-to-fold variation (Supplementary Table S7).

Table 3

Predictive performance for every model, deployment scenario, and outcome, including AUROC and AUPRC with 95% bootstrap confidence intervals, accuracy, sensitivity, specificity, precision, F1-score, and Brier score. Classification metrics were computed at the default probability threshold of 0.50; the same metrics at Youden-optimal and F1-optimal thresholds are provided in Supplementary Table S6.

Scenario	Outcome	Model	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity	Specificity	F1	Brier
Scenario 1	Any latent cardiometabolic disease	LightGBM	0.759 (0.740–0.776)	0.566 (0.535–0.598)	0.653	0.732	0.543	0.194
Scenario 1	Any latent cardiometabolic disease	Logistic regression	0.584 (0.562–0.603)	0.326 (0.301–0.351)	0.551	0.576	0.403	0.244
Scenario 1	Any latent cardiometabolic disease	Random Forest	0.789 (0.773–0.806)	0.634 (0.606–0.662)	0.596	0.821	0.568	0.167
Scenario 1	Any latent cardiometabolic disease	XGBoost	0.653 (0.632–0.671)	0.387 (0.361–0.416)	0.593	0.635	0.454	0.231
Scenario 1	Possible CKD risk	LightGBM	0.874 (0.714–1.000)	0.720 (0.453–1.000)	0.714	0.999	0.667	0.003
Scenario 1	Possible CKD risk	Logistic regression	0.724 (0.597–0.831)	0.008 (0.004–0.020)	0.643	0.728	0.016	0.179
Scenario 1	Possible CKD risk	Random Forest	0.903 (0.794–0.999)	0.669 (0.380–0.910)	0.714	0.998	0.625	0.004
Scenario 1	Possible CKD risk	XGBoost	0.885 (0.767–0.996)	0.375 (0.149–0.688)	0.714	0.96	0.106	0.036
Scenario 1	Undiagnosed diabetes	LightGBM	0.829 (0.792–0.873)	0.575 (0.501–0.647)	0.671	0.894	0.33	0.088
Scenario 1	Undiagnosed diabetes	Logistic regression	0.637 (0.601–0.672)	0.064 (0.051–0.088)	0.665	0.551	0.113	0.239
Scenario 1	Undiagnosed diabetes	Random Forest	0.838 (0.798–0.881)	0.603 (0.530–0.672)	0.636	0.974	0.57	0.048
Scenario 1	Undiagnosed diabetes	XGBoost	0.761 (0.727–0.794)	0.205 (0.150–0.266)	0.711	0.659	0.151	0.197
Scenario 1	Undiagnosed dyslipidemia	LightGBM	0.803 (0.782–0.824)	0.477 (0.432–0.519)	0.66	0.782	0.423	0.16
Scenario 1	Undiagnosed dyslipidemia	Logistic regression	0.639 (0.615–0.663)	0.205 (0.185–0.231)	0.594	0.625	0.289	0.234
Scenario 1	Undiagnosed dyslipidemia	Random Forest	0.823 (0.801–0.844)	0.539 (0.501–0.582)	0.558	0.879	0.471	0.12
Scenario 1	Undiagnosed dyslipidemia	XGBoost	0.698 (0.673–0.721)	0.252 (0.224–0.284)	0.62	0.678	0.328	0.216
Scenario 1	Undiagnosed hypertension	LightGBM	0.765 (0.740–0.788)	0.421 (0.380–0.462)	0.628	0.771	0.39	0.169

Scenario	Outcome	Model	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity	Specificity	F1	Brier
Scenario 1	Undiagnosed hypertension	Logistic regression	0.576 (0.551–0.602)	0.156 (0.140–0.178)	0.569	0.56	0.246	0.244
Scenario 1	Undiagnosed hypertension	Random Forest	0.765 (0.741–0.787)	0.474 (0.432–0.515)	0.478	0.886	0.421	0.129
Scenario 1	Undiagnosed hypertension	XGBoost	0.661 (0.638–0.684)	0.209 (0.185–0.240)	0.659	0.586	0.291	0.225
Scenario 2	Any latent cardiometabolic disease	LightGBM	0.887 (0.874–0.898)	0.763 (0.736–0.788)	0.781	0.849	0.708	0.131
Scenario 2	Any latent cardiometabolic disease	Logistic regression	0.683 (0.665–0.703)	0.406 (0.378–0.438)	0.644	0.637	0.484	0.228
Scenario 2	Any latent cardiometabolic disease	Random Forest	0.913 (0.902–0.922)	0.833 (0.813–0.851)	0.742	0.909	0.743	0.109
Scenario 2	Any latent cardiometabolic disease	XGBoost	0.796 (0.781–0.809)	0.602 (0.567–0.632)	0.683	0.739	0.566	0.185
Scenario 2	Possible CKD risk	LightGBM	0.965 (0.921–1.000)	0.727 (0.465–1.000)	0.714	1.0	0.833	0.001
Scenario 2	Possible CKD risk	Logistic regression	0.844 (0.762–0.914)	0.020 (0.007–0.094)	0.643	0.813	0.023	0.138
Scenario 2	Possible CKD risk	Random Forest	0.945 (0.875–1.000)	0.733 (0.481–1.000)	0.714	1.0	0.833	0.001
Scenario 2	Possible CKD risk	XGBoost	0.948 (0.887–1.000)	0.721 (0.452–1.000)	0.714	0.99	0.303	0.012
Scenario 2	Undiagnosed diabetes	LightGBM	0.871 (0.836–0.907)	0.677 (0.614–0.744)	0.642	0.971	0.558	0.043
Scenario 2	Undiagnosed diabetes	Logistic regression	0.737 (0.700–0.773)	0.143 (0.106–0.200)	0.682	0.613	0.131	0.222
Scenario 2	Undiagnosed diabetes	Random Forest	0.901 (0.872–0.928)	0.705 (0.640–0.772)	0.63	0.999	0.768	0.027
Scenario 2	Undiagnosed diabetes	XGBoost	0.807 (0.775–0.840)	0.324 (0.258–0.391)	0.717	0.743	0.191	0.165
Scenario 2	Undiagnosed dyslipidemia	LightGBM	0.857 (0.836–0.874)	0.607 (0.568–0.645)	0.692	0.843	0.504	0.125
Scenario 2	Undiagnosed dyslipidemia	Logistic regression	0.654 (0.629–0.679)	0.232 (0.206–0.265)	0.601	0.623	0.291	0.231
Scenario 2	Undiagnosed dyslipidemia	Random Forest	0.879 (0.864–0.894)	0.666 (0.630–0.700)	0.516	0.96	0.578	0.088
Scenario 2	Undiagnosed dyslipidemia	XGBoost	0.762 (0.741–	0.326 (0.293–	0.686	0.705	0.375	0.2

Scenario	Outcome	Model	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity	Specificity	F1	Brier
			0.785)	0.366)				
Scenario 2	Undiagnosed hypertension	LightGBM	0.968 (0.962–0.972)	0.795 (0.762–0.827)	0.908	0.905	0.707	0.065
Scenario 2	Undiagnosed hypertension	Logistic regression	0.784 (0.767–0.801)	0.259 (0.236–0.287)	0.747	0.741	0.421	0.197
Scenario 2	Undiagnosed hypertension	Random Forest	0.973 (0.969–0.977)	0.848 (0.823–0.872)	0.846	0.933	0.732	0.056
Scenario 2	Undiagnosed hypertension	XGBoost	0.935 (0.928–0.943)	0.650 (0.608–0.690)	0.973	0.773	0.548	0.119
Scenario 3	Any latent cardiometabolic disease	LightGBM	0.951 (0.944–0.958)	0.865 (0.843–0.886)	0.903	0.878	0.804	0.085
Scenario 3	Any latent cardiometabolic disease	Logistic regression	0.708 (0.691–0.728)	0.435 (0.407–0.468)	0.638	0.681	0.504	0.218
Scenario 3	Any latent cardiometabolic disease	Multi-task MLP	0.864 (0.854–0.875)	0.660 (0.629–0.689)	0.871	0.723	0.658	0.15
Scenario 3	Any latent cardiometabolic disease	Random Forest	0.964 (0.959–0.970)	0.912 (0.896–0.925)	0.907	0.903	0.832	0.076
Scenario 3	Any latent cardiometabolic disease	Residual multi-task MLP	0.921 (0.912–0.929)	0.799 (0.773–0.825)	0.872	0.811	0.726	0.117
Scenario 3	Any latent cardiometabolic disease	XGBoost	0.885 (0.875–0.896)	0.714 (0.685–0.742)	0.856	0.754	0.673	0.141
Scenario 3	Possible CKD risk	LightGBM	0.999 (0.999–1.000)	0.903 (0.767–1.000)	0.714	1.0	0.769	0.001
Scenario 3	Possible CKD risk	Logistic regression	0.917 (0.802–0.980)	0.161 (0.029–0.379)	0.786	0.907	0.055	0.071
Scenario 3	Possible CKD risk	Multi-task MLP	0.951 (0.879–0.999)	0.650 (0.385–0.892)	0.8	0.987	0.304	0.01
Scenario 3	Possible CKD risk	Random Forest	0.999 (0.998–1.000)	0.913 (0.773–1.000)	0.714	1.0	0.833	0.001
Scenario 3	Possible CKD risk	Residual multi-task MLP	0.959 (0.897–1.000)	0.805 (0.577–1.000)	0.800	0.998	0.667	0.002
Scenario 3	Possible CKD risk	XGBoost	0.997 (0.991–1.000)	0.833 (0.630–1.000)	0.786	0.997	0.611	0.004
Scenario 3	Undiagnosed diabetes	LightGBM	0.883 (0.851–0.915)	0.698 (0.634–0.766)	0.642	0.98	0.612	0.033
Scenario 3	Undiagnosed diabetes	Logistic regression	0.753 (0.720–0.787)	0.128 (0.099–0.171)	0.711	0.623	0.139	0.219

Scenario	Outcome	Model	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity	Specificity	F1	Brier
Scenario 3	Undiagnosed diabetes	Multi-task MLP	0.829 (0.800–0.855)	0.300 (0.241–0.366)	0.772	0.731	0.218	0.162
Scenario 3	Undiagnosed diabetes	Random Forest	0.922 (0.898–0.943)	0.729 (0.668–0.790)	0.618	1.0	0.764	0.023
Scenario 3	Undiagnosed diabetes	Residual multi-task MLP	0.878 (0.848–0.905)	0.601 (0.533–0.668)	0.706	0.886	0.357	0.086
Scenario 3	Undiagnosed diabetes	XGBoost	0.845 (0.817–0.872)	0.382 (0.317–0.449)	0.74	0.767	0.211	0.148
Scenario 3	Undiagnosed dyslipidemia	LightGBM	0.984 (0.981–0.987)	0.896 (0.870–0.916)	0.943	0.942	0.81	0.042
Scenario 3	Undiagnosed dyslipidemia	Logistic regression	0.788 (0.767–0.810)	0.342 (0.305–0.379)	0.726	0.724	0.406	0.19
Scenario 3	Undiagnosed dyslipidemia	Multi-task MLP	0.942 (0.934–0.949)	0.653 (0.611–0.695)	0.964	0.806	0.573	0.11
Scenario 3	Undiagnosed dyslipidemia	Random Forest	0.984 (0.981–0.987)	0.897 (0.876–0.916)	0.955	0.939	0.807	0.043
Scenario 3	Undiagnosed dyslipidemia	Residual multi-task MLP	0.966 (0.960–0.972)	0.792 (0.754–0.829)	0.964	0.863	0.652	0.088
Scenario 3	Undiagnosed dyslipidemia	XGBoost	0.966 (0.961–0.971)	0.779 (0.741–0.810)	0.998	0.834	0.642	0.087
Scenario 3	Undiagnosed hypertension	LightGBM	0.972 (0.967–0.976)	0.804 (0.769–0.838)	0.901	0.924	0.741	0.057
Scenario 3	Undiagnosed hypertension	Logistic regression	0.786 (0.768–0.805)	0.267 (0.243–0.296)	0.745	0.733	0.414	0.195
Scenario 3	Undiagnosed hypertension	Multi-task MLP	0.911 (0.901–0.921)	0.551 (0.506–0.595)	0.944	0.745	0.514	0.141
Scenario 3	Undiagnosed hypertension	Random Forest	0.977 (0.974–0.981)	0.864 (0.843–0.887)	0.852	0.944	0.761	0.051
Scenario 3	Undiagnosed hypertension	Residual multi-task MLP	0.937 (0.928–0.946)	0.671 (0.632–0.715)	0.868	0.851	0.602	0.098
Scenario 3	Undiagnosed hypertension	XGBoost	0.944 (0.937–0.951)	0.661 (0.618–0.703)	0.982	0.792	0.573	0.112

Discrimination and precision-recall analysis

Receiver operating characteristic curves for all six models under Scenario 3 are shown in Fig. 5, and precision-recall curves in Fig. 6. Precision-recall analysis, which is more sensitive to performance on the diseased minority class, reinforced the discrimination findings. Tree-ensemble models achieved substantially higher average precision than logistic regression across all outcomes, with the largest margins for dyslipidemia and hypertension. Because screening operates under class imbalance, this advantage translates into more efficient identification of true positives at a given alert rate.

Calibration

Calibration curves under Scenario 3 are presented in Fig. 7, and Brier scores are reported in Table 3. Agreement between predicted and observed risk was generally better for the more prevalent outcomes, and Brier scores were lowest for the high-performing tree ensembles under Scenario 3, for example 0.076 for Random Forest compared with 0.218 for logistic regression on the composite outcome. These differences in native calibration motivate the conformal analysis, which provides coverage guarantees regardless of the base model's calibration.

Clinical utility

Decision-curve analysis (Fig. 8) showed that, relative to treat-all and treat-none strategies, the tree-ensemble models generally provided the greatest net benefit across the range of clinically relevant threshold probabilities, particularly for outcomes with strong discrimination under Scenario 3. The discrimination advantages of the nonlinear models therefore correspond to genuine decision-support value rather than to a metric artifact.

Learned latent representations

A two-dimensional UMAP projection of the 128-dimensional embeddings learned by the residual multi-task network is shown in Fig. 9. The projection displayed a continuous gradient related to disease burden rather than a set of discrete, well-separated clusters. Consistent with this, the silhouette coefficient of the full 128-dimensional embedding, computed with the number of co-occurring conditions as the grouping label, was close to zero (-0.02 by disease count, and 0.07 for any latent condition versus none; Supplementary Figure S1), indicating that the shared encoder did not partition patients into distinct clusters but instead encoded disease burden as a continuum. This continuous structure is expected, because cardiometabolic abnormalities accumulate along a spectrum rather than forming discrete phenotypes, and it is consistent with the multi-task objective, which encourages a shared representation that is simultaneously informative for all five outcomes.

Uncertainty-aware screening with conformal prediction

Conformal prediction results across all six models and five outcomes under Scenario 3 are summarized in Fig. 10 and Table 4. Empirical coverage was close to the 90% target for every model and outcome, in the range of about 0.88 to 0.92, which confirms that the distribution-free coverage guarantee held empirically, including for models with poor native calibration.

The main finding is that models with similar coverage differed markedly in the certainty they could support. For the composite outcome, the uncertainty rate, that is the fraction of individuals assigned an ambiguous prediction set and therefore flagged for confirmatory testing, was 0.3% for Random Forest and 4.7% for LightGBM, 15.1% for the residual multi-task network, and 59.9% for logistic regression, with values of 26.4% for XGBoost and 30.9% for the multi-task network. At a guaranteed 90% coverage, logistic regression can confidently classify fewer than half of individuals for the composite outcome, whereas the strongest tree ensembles confidently classify nearly all of them, a difference that AUROC does not reveal. A similar pattern held across the individual outcomes; for undiagnosed hypertension, the uncertainty rate was 0.0% for LightGBM and Random Forest compared with 53.2% for logistic regression. For the very rare CKD-risk outcome, conformal sets were frequently singletons or empty, with an empty rate near 0.10 across models, which is an appropriate consequence of a near-zero prevalence.

In operational terms, the uncertainty rate defines the referral workload of a screening deployment. A low uncertainty rate at guaranteed coverage means that most individuals receive a trustworthy screen-positive or screen-negative result, while the minority in the ambiguous set are triaged to confirmatory testing. This behavior is what a responsible population-screening system should show, and it cannot be read from discrimination metrics alone.

Table 4

Conformal prediction summary for every model and outcome under the routine clinical assessment scenario, including target and empirical coverage, average prediction set size, uncertainty rate, singleton rate, and empty rate.

Model	Outcome	Target coverage	Empirical coverage	Average set size	Uncertainty rate	Singleton rate	Empty rate	Calibration N	Evaluation N
LightGBM	Any latent cardiometabolic disease	0.9	0.909	1.047	0.047	0.953	0.0	2037	2038
Logistic regression	Any latent cardiometabolic disease	0.9	0.899	1.599	0.599	0.401	0.0	2037	2038
Multi-task MLP	Any latent cardiometabolic disease	0.9	0.895	1.309	0.309	0.691	0.0	2037	2038
Random Forest	Any latent cardiometabolic disease	0.9	0.908	1.003	0.003	0.997	0.0	2037	2038
Residual multi-task MLP	Any latent cardiometabolic disease	0.9	0.899	1.151	0.151	0.849	0.0	2037	2038
XGBoost	Any latent cardiometabolic disease	0.9	0.899	1.264	0.264	0.736	0.0	2037	2038
LightGBM	Possible CKD risk	0.9	0.895	0.895	0.0	0.895	0.105	2037	2038
Logistic regression	Possible CKD risk	0.9	0.906	0.994	0.0	0.994	0.006	2037	2038
Multi-task MLP	Possible CKD risk	0.9	0.886	0.887	0.0	0.887	0.113	2037	2038
Random Forest	Possible CKD risk	0.9	0.903	0.903	0.0	0.903	0.097	2037	2038
Residual multi-task MLP	Possible CKD risk	0.9	0.892	0.893	0.0	0.893	0.107	2037	2038
XGBoost	Possible CKD risk	0.9	0.899	0.899	0.0	0.899	0.101	2037	2038
LightGBM	Undiagnosed diabetes	0.9	0.894	0.91	0.0	0.91	0.09	2037	2038
Logistic regression	Undiagnosed diabetes	0.9	0.895	1.576	0.576	0.424	0.0	2037	2038
Multi-task MLP	Undiagnosed diabetes	0.9	0.898	1.359	0.359	0.641	0.0	2037	2038
Random Forest	Undiagnosed diabetes	0.9	0.884	0.895	0.0	0.895	0.105	2037	2038
Residual multi-task MLP	Undiagnosed diabetes	0.9	0.897	1.047	0.047	0.953	0.0	2037	2038
XGBoost	Undiagnosed diabetes	0.9	0.89	1.274	0.274	0.726	0.0	2037	2038
LightGBM	Undiagnosed dyslipidemia	0.9	0.905	0.939	0.0	0.939	0.061	2037	2038
Logistic regression	Undiagnosed dyslipidemia	0.9	0.897	1.44	0.44	0.56	0.0	2037	2038
Multi-task MLP	Undiagnosed dyslipidemia	0.9	0.901	1.132	0.132	0.868	0.0	2037	2038

Model	Outcome	Target coverage	Empirical coverage	Average set size	Uncertainty rate	Singleton rate	Empty rate	Calibration N	Evaluation N
Random Forest	Undiagnosed dyslipidemia	0.9	0.898	0.932	0.0	0.932	0.068	2037	2038
Residual multi-task MLP	Undiagnosed dyslipidemia	0.9	0.895	1.048	0.048	0.952	0.0	2037	2038
XGBoost	Undiagnosed dyslipidemia	0.9	0.903	1.068	0.068	0.932	0.0	2037	2038
LightGBM	Undiagnosed hypertension	0.9	0.903	0.97	0.0	0.97	0.03	2037	2038
Logistic regression	Undiagnosed hypertension	0.9	0.913	1.532	0.532	0.468	0.0	2037	2038
Multi-task MLP	Undiagnosed hypertension	0.9	0.915	1.264	0.264	0.736	0.0	2037	2038
Random Forest	Undiagnosed hypertension	0.9	0.909	0.95	0.0	0.95	0.05	2037	2038
Residual multi-task MLP	Undiagnosed hypertension	0.9	0.913	1.115	0.115	0.885	0.0	2037	2038
XGBoost	Undiagnosed hypertension	0.9	0.904	1.159	0.159	0.841	0.0	2037	2038

Discussion

Using NHANES 2011 to 2014, we developed and evaluated a unified screening framework for five undiagnosed cardiometabolic outcomes across three realistic deployment tiers, with an explicit uncertainty layer. Three findings stand out. Undiagnosed cardiometabolic disease was common, with about one in four adults having at least one previously unrecognized condition, which confirms a large screening gap. Discrimination improved as clinical information increased, and tree-ensemble models reached an AUROC of 0.96 or higher for several outcomes under routine clinical assessment while consistently outperforming logistic regression. Conformal prediction achieved its guaranteed coverage across all models yet revealed large differences in the certainty that each model could support, which is information that discrimination metrics conceal but that is decisive for real-world screening.

The nested-scenario design quantifies a practical trade-off. Richer information yields higher discrimination but at greater cost and lower accessibility. Survey-only models already provided clinically meaningful discrimination for several outcomes, which supports their use in low-resource community pre-screening to prioritize who should proceed to examination or laboratory testing. The largest gains from adding examination and laboratory data occurred for hypertension and dyslipidemia, consistent with the physiological proximity of the added variables to those conditions. Because defining biomarkers were excluded, these gains reflect genuine screening signal rather than leakage.

A screening model that is accurate on average can still be overconfident on individual atypical patients. Conformal prediction addresses this by attaching a coverage-guaranteed prediction set to each individual and by making deferral an explicit output. Our results show that two models with near-identical AUROC can imply very different clinical workflows. At a guaranteed 90% coverage, one model may confidently classify almost all individuals while another defers on the majority. Reporting the uncertainty rate therefore gives decision-makers a direct measure of referral volume and of where human review is most needed, which links predictive modeling to clinical deployment in a way that ranking metrics do not.

Our study extends a growing body of NHANES-based cardiometabolic prediction research. Earlier work developed survey-design-aware neural network risk scores with uncertainty quantification for diabetes in the same NHANES window and showed that models of varying complexity yield generalizable population-level results [30]. We build on that direction by moving from a single disease to a multi-outcome framework with a shared-encoder network, by structuring the analysis around cost-tiered deployment scenarios, and by using conformal prediction to produce an operational referral signal. Our observation that tree ensembles are strong discriminators on tabular clinical data is consistent with the broader literature on tabular machine learning [16, 17], and single-disease NHANES models have reported comparable discrimination without an uncertainty layer [19, 29]. The margin by which tree ensembles outperformed logistic regression

here is larger than the small average benefit reported in systematic comparisons of machine learning and regression for clinical prediction [20], which suggests genuine nonlinear structure and interactions in these screening tasks rather than a modeling artifact. At the same time, that same body of evidence cautions against equating higher discrimination with clinical readiness, and our emphasis on calibrated uncertainty is aligned with the wider trustworthy artificial intelligence agenda in medicine [18, 23, 24].

Our prevalence findings are also consistent with national surveillance. The survey-weighted burden of undiagnosed disease we estimated echoes long-standing evidence that a meaningful fraction of diabetes and hypertension in the United States is undiagnosed and that CKD is often unrecognized [7–9]. This concordance supports the external plausibility of the cohort and reinforces that population screening retains substantial unrealized yield.

Equity deserves particular attention before any deployment. Because conformal coverage guarantees are marginal, the coverage and the uncertainty rate could still differ across demographic groups, and prior work has shown how opaque clinical algorithms can disadvantage already underserved populations [31]. We therefore regard subgroup-conditional, or Mondrian, conformal analysis not as an optional extension but as a prerequisite for responsible use, so that the referral burden created by uncertain predictions is distributed fairly across sex, race and ethnicity, age, income, and body mass index.

Translating uncertainty-aware models into practice will also require familiarity with conformal methods among clinical teams, which is still limited despite growing interest [28]. Presenting the output as an interpretable three-way decision, namely confident low risk, confident high risk, or refer for confirmatory testing, may ease adoption more than continuous risk scores that clinicians must threshold themselves, and it makes the model's decision to abstain visible and auditable.

The framework maps onto a staged screening pathway. A low-cost survey-only model can support community pre-screening, models with examination and laboratory data can be used where warranted, and a conformal uncertainty layer at every tier separates confident screen-negative and screen-positive individuals from an uncertain group routed to confirmatory testing. Such a design could improve the yield of population screening while limiting the burden of unnecessary follow-up, and it provides an explicit and auditable place for clinician oversight.

Strengths and limitations

This study has several strengths. It used a large, nationally representative sample with objective biomarkers, a principled definition of undiagnosed disease combined with a strict rule against diagnostic leakage, a broad model comparison spanning linear, ensemble, and neural network families, a deployment-tiered evaluation, and a distribution-free uncertainty layer with empirically verified coverage. The multi-task architecture additionally produced a shared latent representation spanning the cardiometabolic spectrum.

Several limitations should be considered. First, NHANES is cross-sectional, so the outcomes reflect prevalent undiagnosed disease at a single point rather than incident disease, and causal or prognostic claims are not warranted. Second, possible CKD risk was defined by albuminuria alone, without the estimated glomerular filtration rate or confirmation of chronicity, and its very low prevalence makes its precision estimates unstable despite a high AUROC. Third, although population prevalence was estimated with the complex-survey design, the predictive models themselves were fitted on unweighted data and used median imputation; survey-weighted model estimation and multiply-imputed sensitivity analyses are planned to further confirm generalizability. Fourth, evaluation used a single held-out split from the 2011 to 2014 window, and external validation on later NHANES cycles is needed to assess transportability. Fifth, model interpretation using variable-attribution methods and formal subgroup fairness analyses of coverage and uncertainty across sex, race and ethnicity, age, income, and body mass index are reserved for a dedicated follow-up analysis. Finally, conformal coverage guarantees are marginal, and group-conditional conformal analysis will be used to verify coverage within clinically important subgroups.

Recommendations

Several practical recommendations follow from these findings. First, screening programs should adopt a staged, resource-aware pathway. Survey-only models can support community pre-screening to prioritize individuals for further assessment, examination-based models can be applied in primary care, and laboratory-supported models can be reserved for clinical settings, matching the intensity and cost of testing to the information already available [11, 12]. Second, screening models should be evaluated and reported for calibration and uncertainty, not discrimination alone, because two models with similar AUROC can imply very different referral workloads once coverage is guaranteed [20–22]. Reporting the uncertainty rate alongside AUROC and AUPRC would give decision-makers a transparent estimate of the confirmatory-testing burden that a model implies. Third, model outputs should be presented to clinicians as an interpretable three-way decision, namely confident low risk, confident high risk, or refer for confirmatory testing, rather than as a bare probability, so that abstention becomes an explicit and auditable option [23, 24, 28]. Fourth, before deployment, uncertainty-aware tools

should be audited for equity using subgroup-conditional coverage, so that the burden of uncertain predictions does not fall disproportionately on already underserved groups [31]. Fifth, prospective and external validation on independent cohorts and later survey cycles should precede clinical use, and models should be re-calibrated when they are applied to populations that differ from the development sample [35]. Finally, health systems should integrate these tools into existing decision-support and referral workflows with clear governance over how uncertain cases are escalated, so that any improvement in screening yield is not offset by an unmanaged increase in follow-up testing.

Conclusions

In a nationally representative United States sample, a unified screening framework identified multiple undiagnosed cardiometabolic conditions across community, primary-care, and clinical deployment tiers. Discrimination improved as clinical information increased and was strongest for tree-ensemble models, while a shared-encoder neural network learned a shared latent representation that encoded cardiometabolic disease burden as a continuum. Split conformal prediction attached coverage-guaranteed uncertainty to every prediction and exposed large and actionable differences in the certainty that similarly discriminating models can support. By reporting both what a model predicts and when it should defer, this approach provides a practical foundation for population-level cardiometabolic screening and clinical decision support. Future work will add model interpretation, external temporal validation, survey-weighted model estimation, and subgroup-aware conformal prediction to test whether the uncertainty layer is equitable across demographic and clinical groups.

Abbreviations

AUPRC

area under the precision-recall curve

AUROC

area under the receiver operating characteristic curve

BMI

body mass index

CI

confidence interval

CKD

chronic kidney disease

HbA1c

glycated haemoglobin

HDL

high-density lipoprotein

IQR

interquartile range

MEC

mobile examination centre

MLP

multilayer perceptron

NCHS

National Center for Health Statistics

NHANES

National Health and Nutrition Examination Survey

PSU

primary sampling unit

SD

standard deviation

UMAP

uniform manifold approximation and projection.

Declarations

Board, and all participants provided written informed consent at the time of data collection. No additional ethical approval was required for this secondary analysis.

Ethics approval and consent to participate

This study used publicly available, de-identified secondary data from NHANES. The NHANES protocol was approved by the National Center for Health Statistics Research Ethics Review Board, and all participants provided written informed consent at the time of data collection. No additional ethical approval was required for this secondary analysis.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

KSAN conceived and planned the study. AH and ARR conducted the analysis. KSAN, ARR, AH, and ASA drafted the manuscript. KSAN, KMAAS, SA, ARR, and AH critically reviewed and approved the final manuscript.

Acknowledgement

The authors thank the National Center for Health Statistics and the NHANES participants for the public availability of these data.

Data Availability

The NHANES 2011 to 2014 datasets analyzed in this study are publicly available from the United States Centers for Disease Control and Prevention at <https://www.cdc.gov/nchs/nhanes/>. The analysis code and derived model-ready datasets generated during this study are available from the corresponding author on reasonable request.

References

1. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10258):1223–49.
2. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol*. 2020;76(25):2982–3021.
3. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract*. 2022;183:109119.
4. Fang M, Wang D, Coresh J, Selvin E. Trends in diabetes treatment and control in U.S. adults, 1999–2018. *N Engl J Med*. 2021;384(23):2219–28.
5. Muntner P, Hardy ST, Fine LJ, Jaeger BC, Wozniak G, Levitan EB, et al. Trends in blood pressure control among US adults with hypertension, 1999–2000 to 2017–2018. *JAMA*. 2020;324(12):1190–200.
6. Murphy D, McCulloch CE, Lin F, Banerjee T, Bragg-Gresham JL, Eberhardt MS, et al. Trends in prevalence of chronic kidney disease in the United States. *Ann Intern Med*. 2016;165(7):473–81.
7. Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. *JAMA*. 2015;314(10):1021–9.

8. Bikbov B, Purcell CA, Levey AS, Smith M, Abdoli A, Abebe M, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2020;395(10225):709–33.
9. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH et al. /ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. Hypertension. 2018;71(6):e13–e115.
10. Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, et al. Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet*. 2010;375(9733):2215–22.
11. Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: World Health Organization; 1968.
12. US Preventive Services Task Force. Screening for prediabetes and type 2 diabetes: US Preventive Services Task Force recommendation statement. *JAMA*. 2021;326(8):736–43.
13. Centers for Disease Control and Prevention, National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES). Hyattsville, MD: NCHS. Available from: <https://www.cdc.gov/nchs/nhanes/>
14. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–58.
15. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–30.
16. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Inf Process Syst*. 2022;35:507–20.
17. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Inf Fusion*. 2022;81:84–90.
18. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
19. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. *Comput Struct Biotechnol J*. 2017;15:104–16.
20. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
21. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. *Proc Int Conf Mach Learn*. 2017;70:1321–30.
22. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
23. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*. 2019;1(1):20–3.
24. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
25. Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. New York: Springer; 2005.
26. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511 [Preprint]. 2021.
27. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. *J Am Stat Assoc*. 2018;113(523):1094–111.
28. Vazquez J, Facelli JC. Conformal prediction in clinical medical sciences. *J Healthc Inf Res*. 2022;6(3):241–52.
29. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inf Decis Mak*. 2019;19(1):211.
30. Matabuena M, Vidal JC, Ghosal R, Onnela JP. Screening for diabetes mellitus in the US population using neural network-based modelling and complex survey designs. *Stat Methods Med Res*. 2026 (online ahead of print). Preprint available at arXiv:2403.19752.
31. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–53.
32. Romano Y, Sesia M, Candès EJ. Classification with valid and adaptive coverage. *Adv Neural Inf Process Syst*. 2020;33:3581–91.
33. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–74.
34. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv:1802.03426 [Preprint]. 2018.
35. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.

Figures

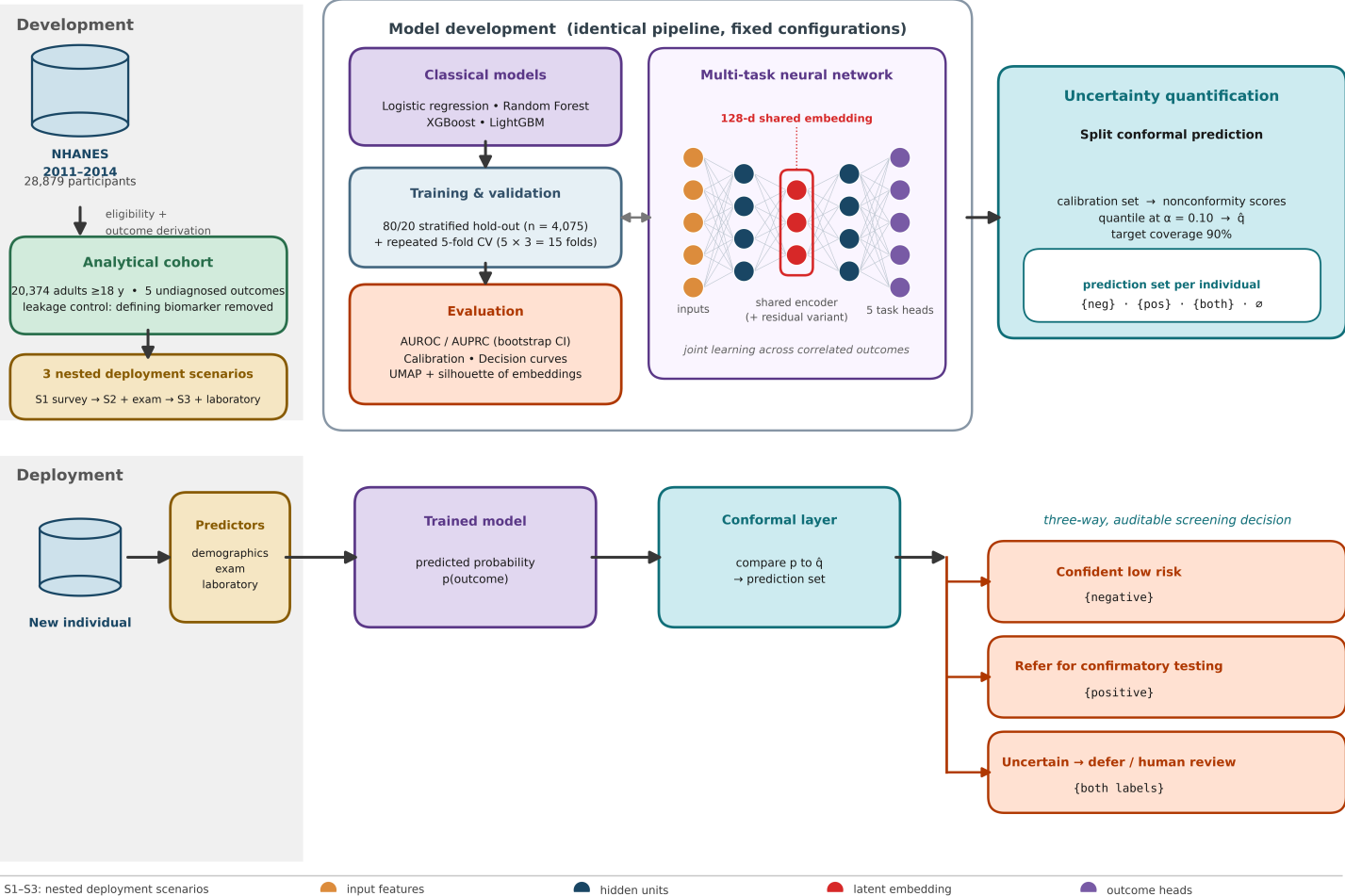


Figure 1

Methodological framework of the study. The end-to-end pipeline proceeds from NHANES 2011 to 2014 data integration and cohort assembly, through the definition of five undiagnosed cardiometabolic outcomes with strict leakage control and three nested deployment scenarios of increasing clinical information, to the development of six predictive models under an identical pipeline with repeated cross-validation, multi-faceted evaluation (discrimination, calibration, clinical utility, and learned representations), and split conformal prediction that converts each model output into a three-way, uncertainty-aware screening decision: confident low risk, refer for confirmatory testing, or defer for review.

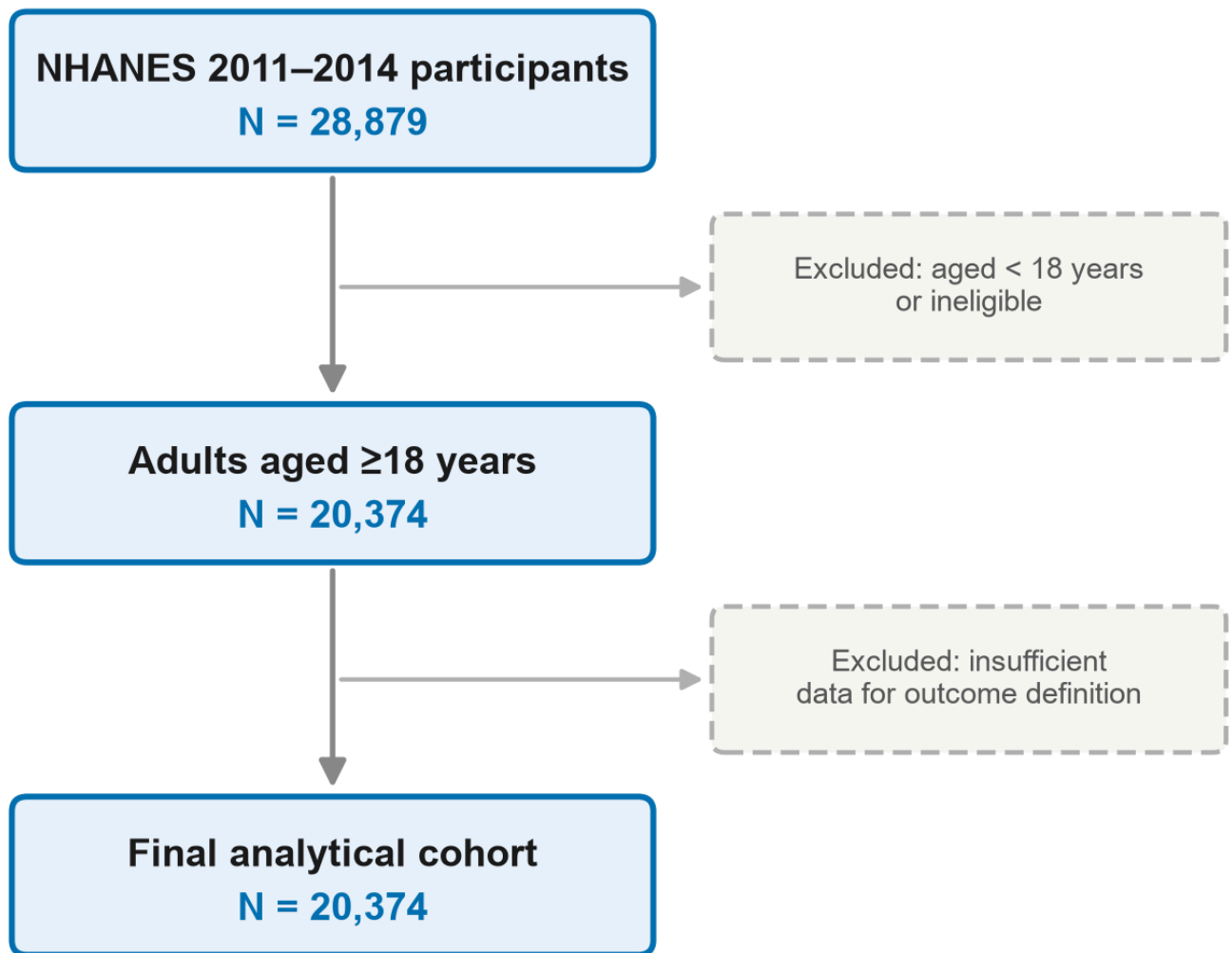


Figure 2

Study population flowchart. Of 28,879 NHANES 2011 to 2014 participants, 20,374 adults aged 18 years or older with sufficient information for outcome definition formed the analytical cohort.

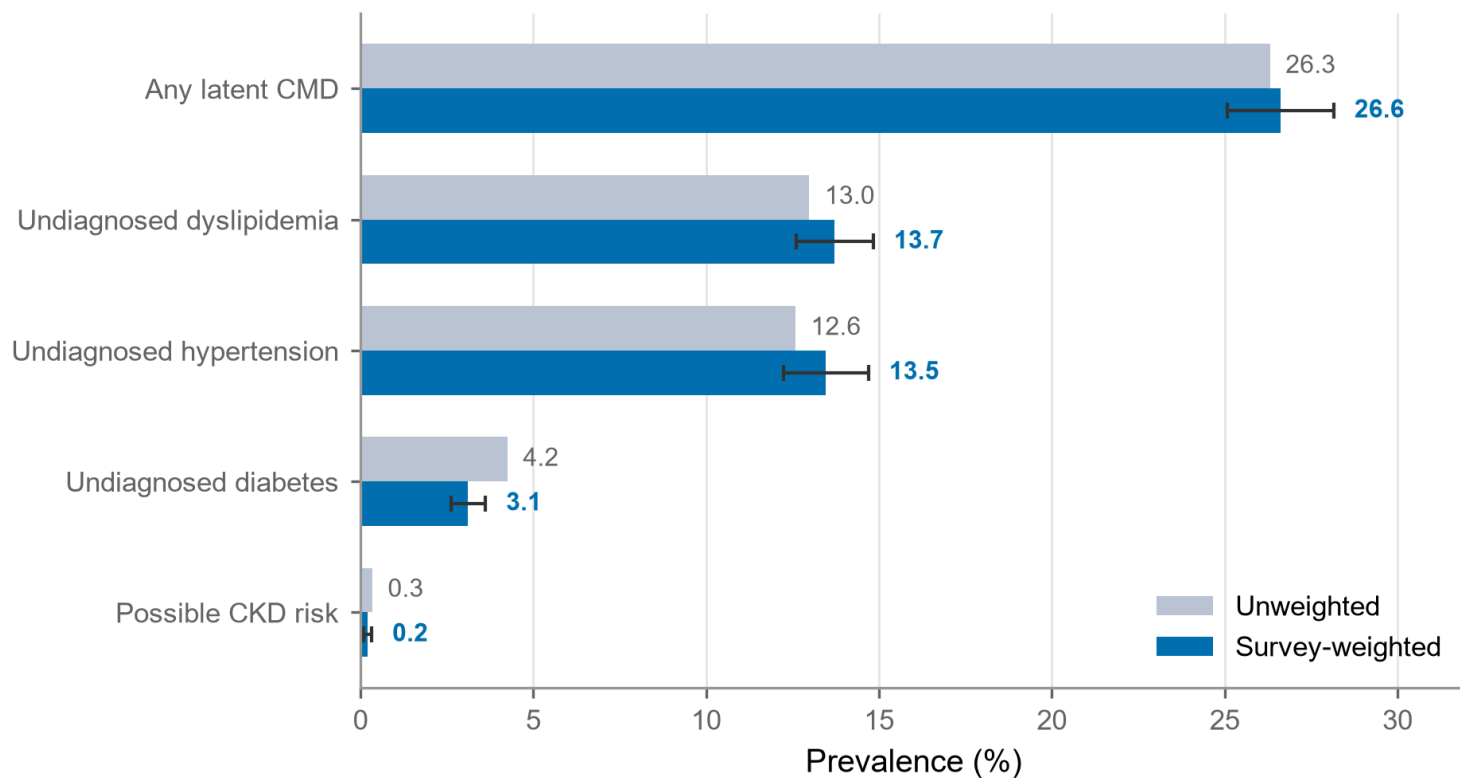


Figure 3

Prevalence of the five undiagnosed cardiometabolic outcomes in the analytical cohort (N = 20,374), showing unweighted proportions and survey-weighted estimates with 95% confidence intervals.

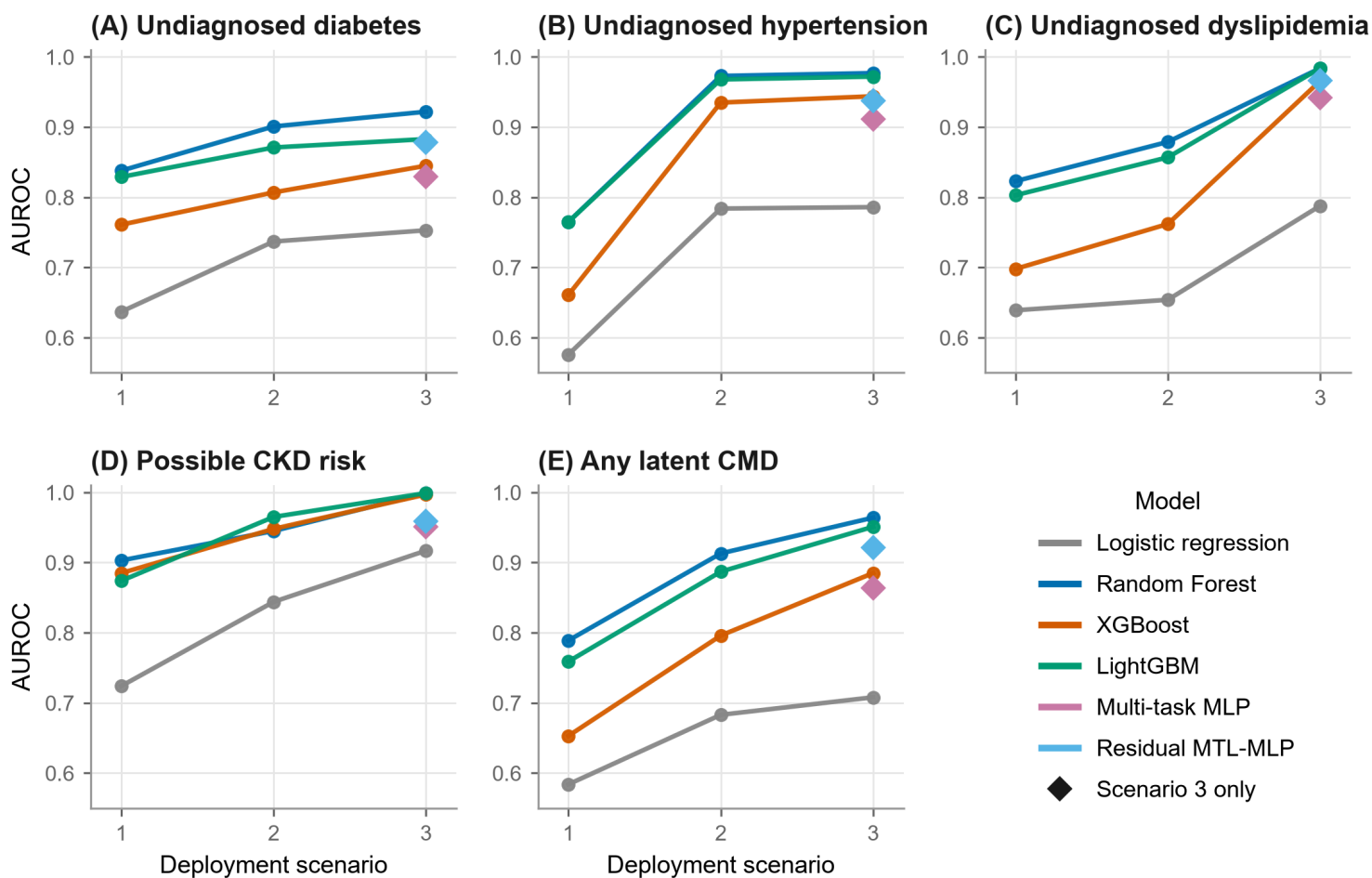


Figure 4

Model discrimination (AUROC) across the three deployment scenarios and all outcomes, for every model.

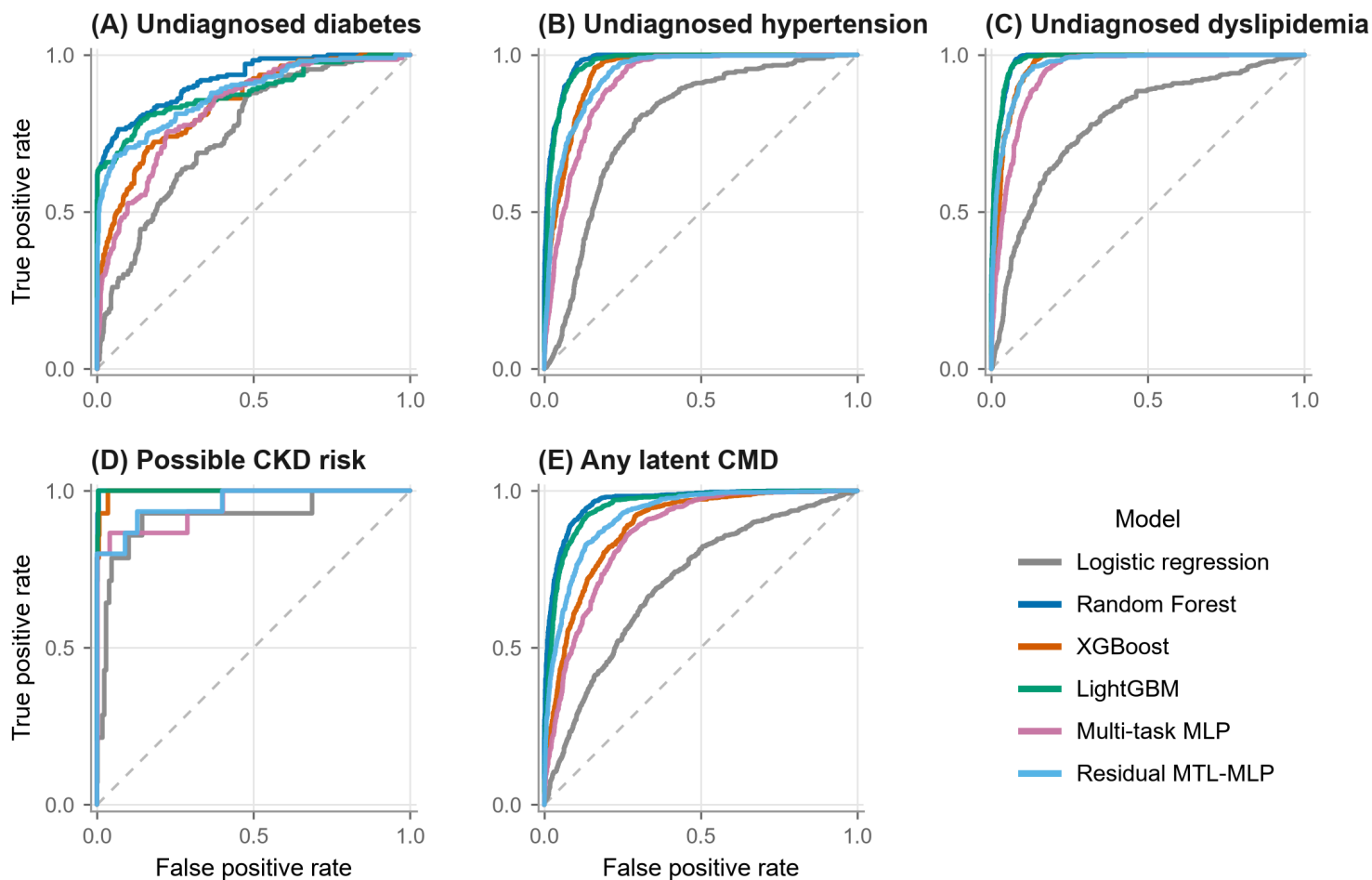


Figure 5

Receiver operating characteristic curves comparing all six models under the routine clinical assessment scenario.

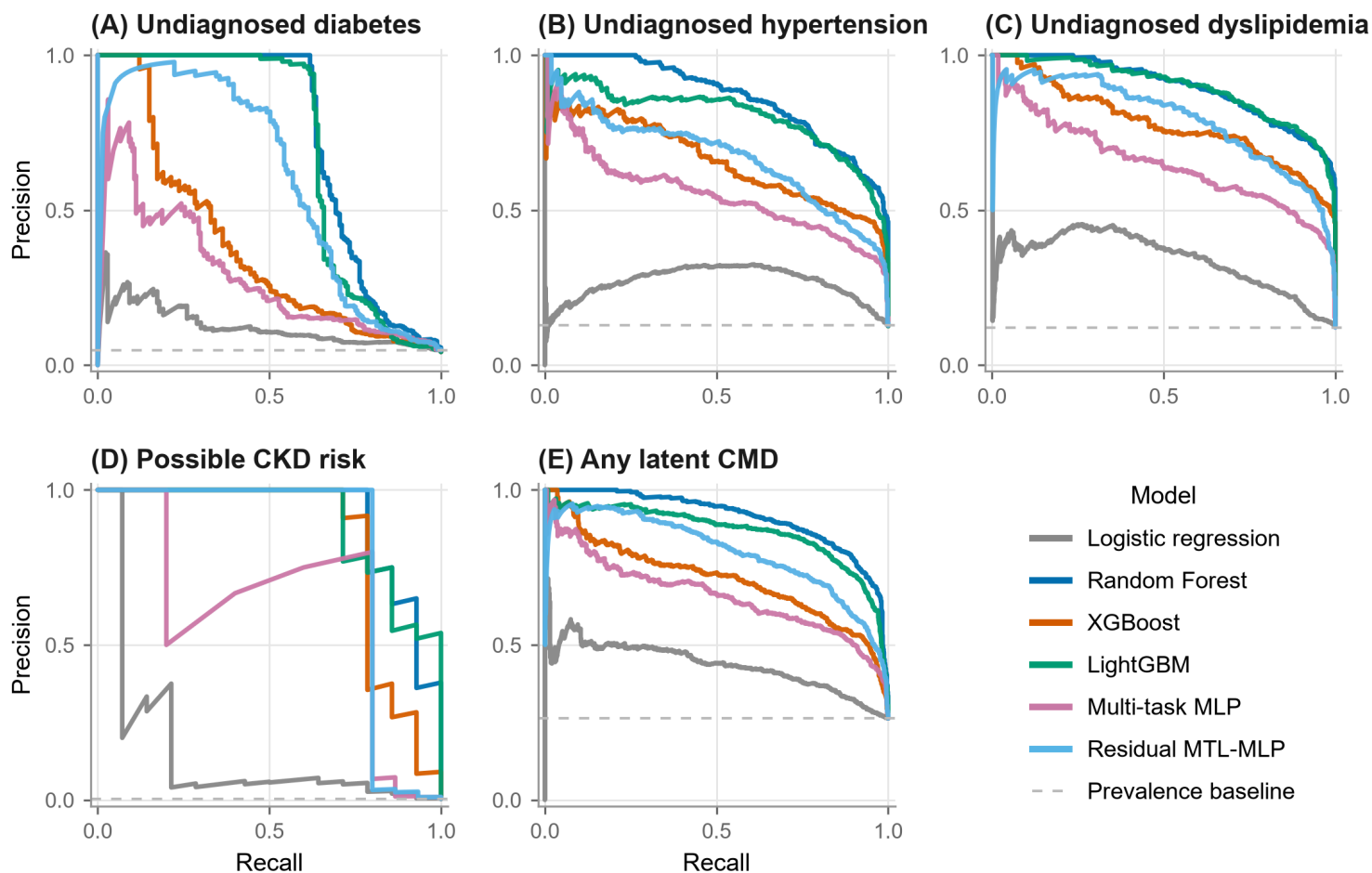


Figure 6

Precision-recall curves comparing all six models under the routine clinical assessment scenario.

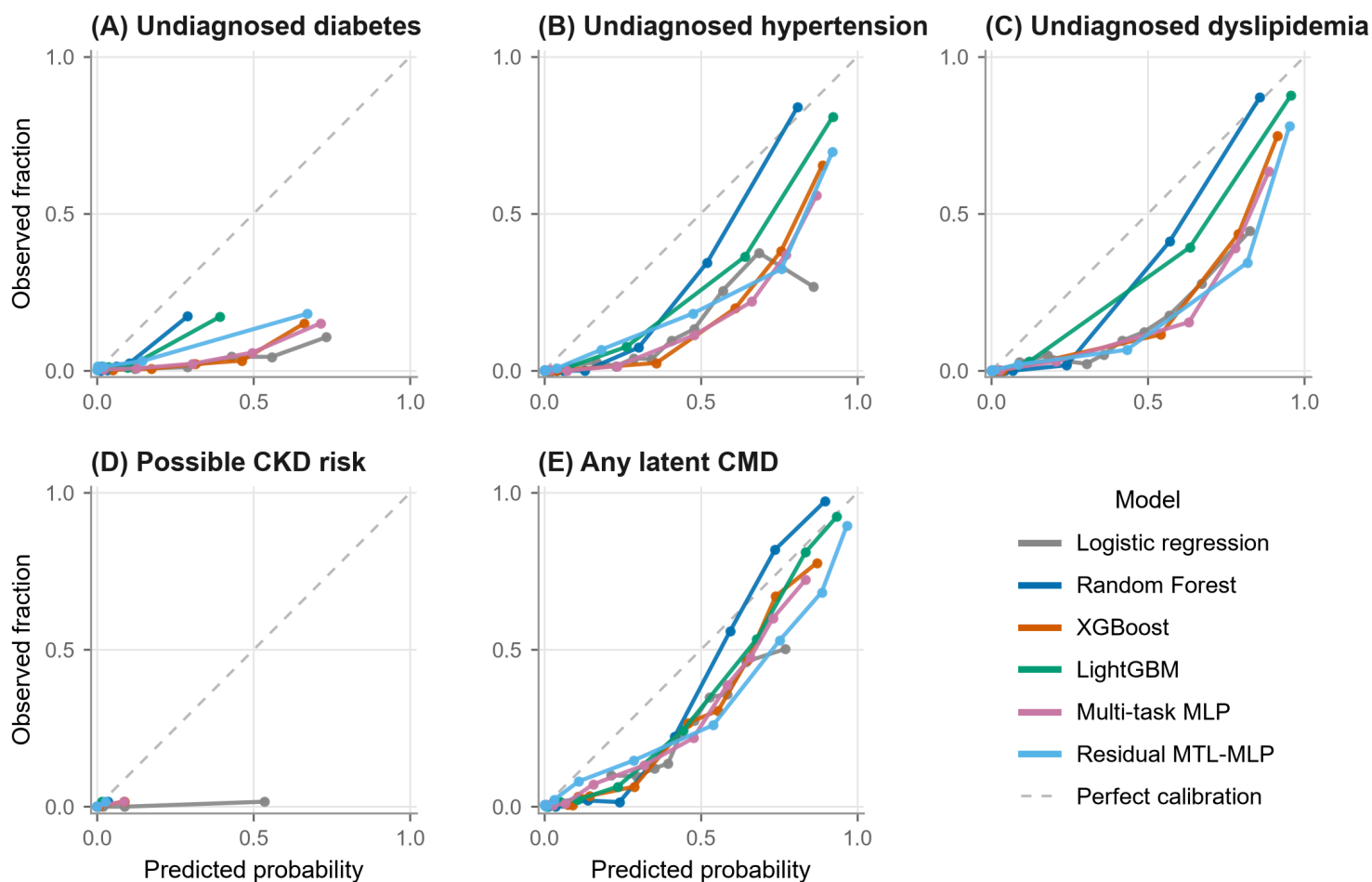


Figure 7

Calibration curves under the routine clinical assessment scenario.

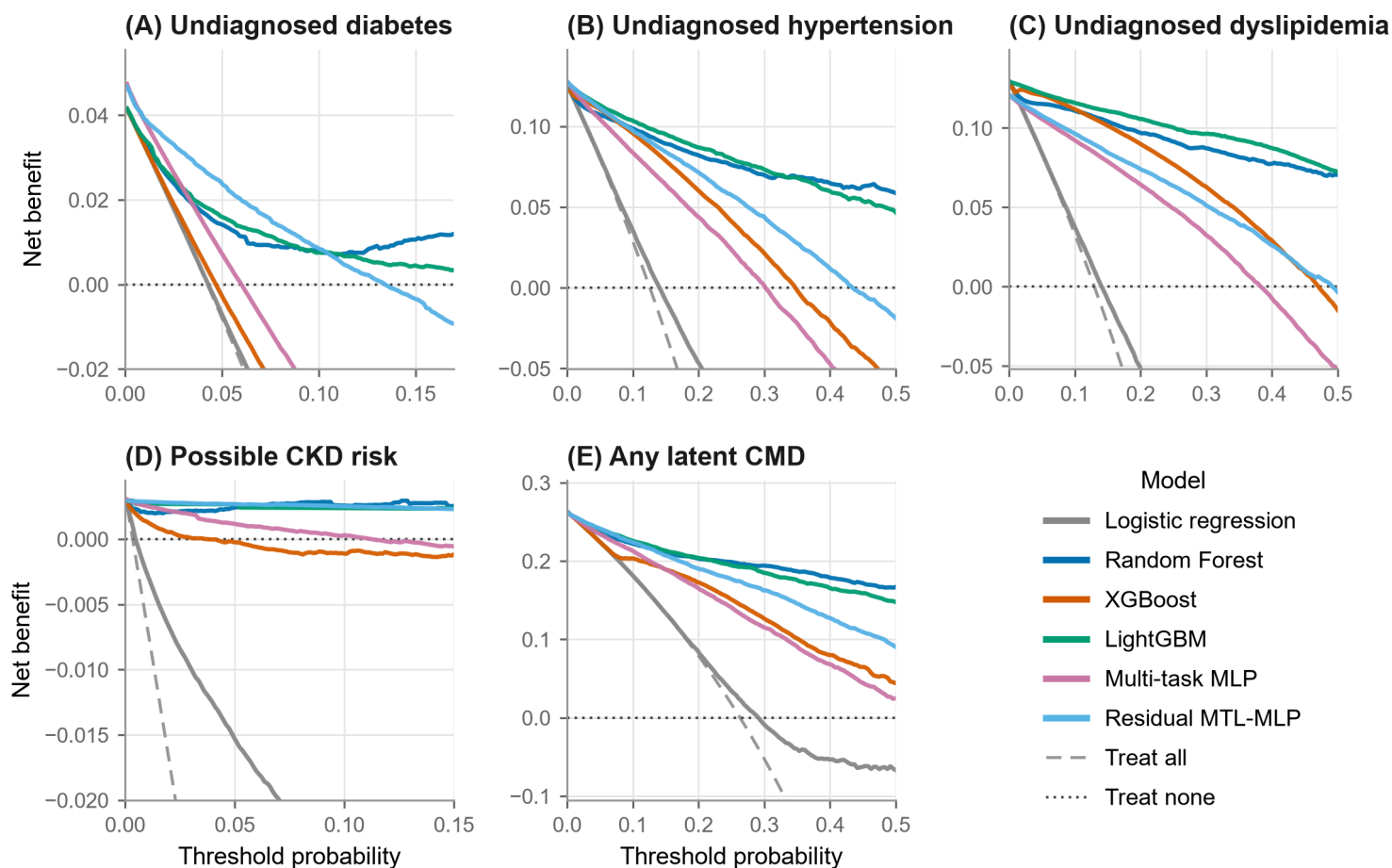


Figure 8

Decision-curve analysis showing net benefit relative to treat-all and treat-none strategies across threshold probabilities under the routine clinical assessment scenario.

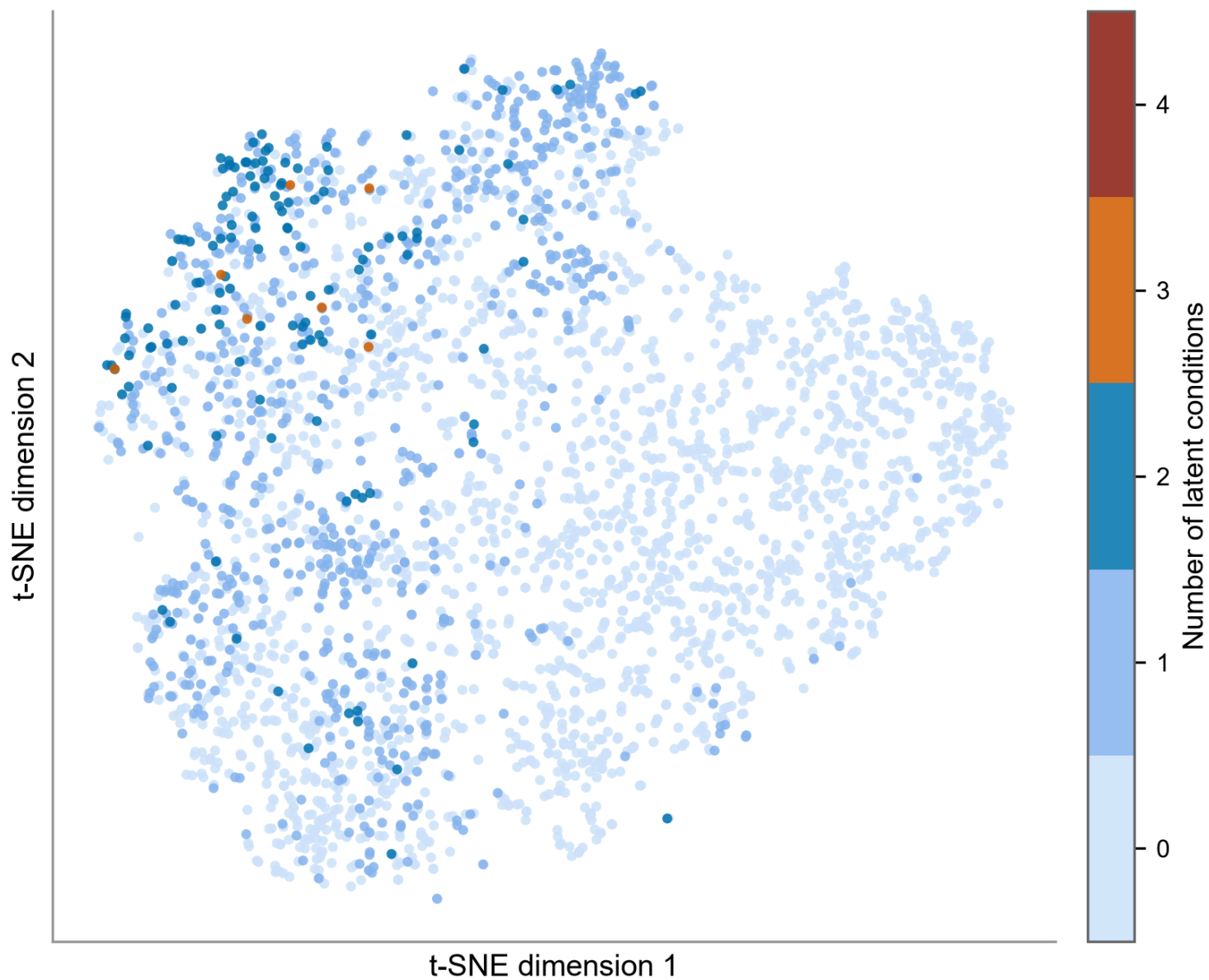


Figure 9

Two-dimensional UMAP projection of the 128-dimensional patient embeddings learned by the residual multi-task neural network, coloured by the number of co-occurring undiagnosed conditions. The projection shows a continuous gradient in disease burden rather than discrete clusters; the silhouette coefficient of the full 128-dimensional embedding was close to zero (Supplementary Figure S1).

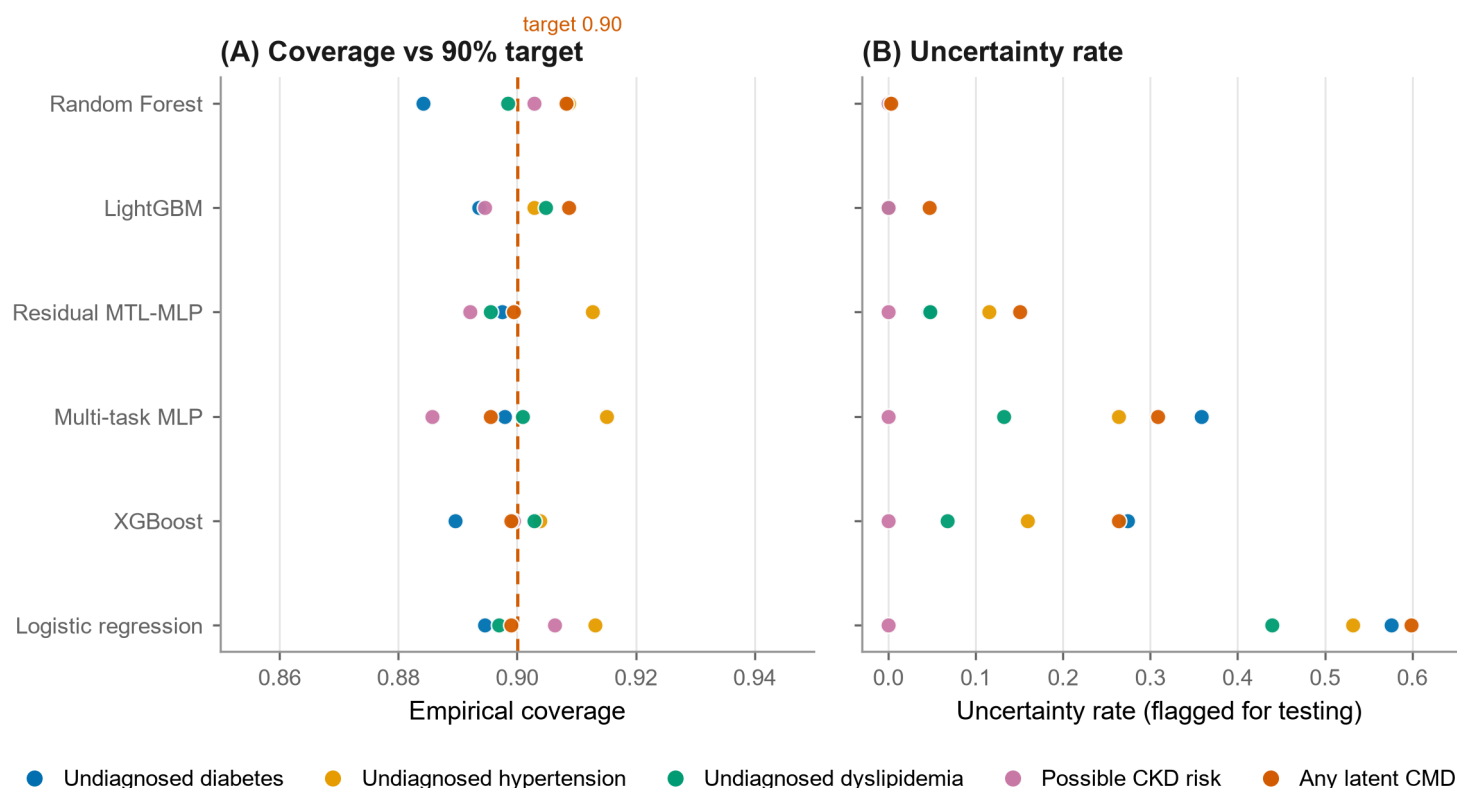


Figure 10

Conformal prediction performance across all models under the routine clinical assessment scenario. (A) Empirical coverage for each model and outcome relative to the 90% target; (B) the uncertainty rate, that is the proportion of individuals assigned an ambiguous prediction set and therefore flagged for confirmatory testing.

Supplementary Files

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

- [SupplementaryMaterials.docx](#)
- [figureS1silhouette.png](#)