Clinical utility gains from incorporating comorbidity and geographic location information into risk estimation equations for atherosclerotic cardiovascular disease

Yizhe Xu\textsuperscript{1}, Agata Foryciarz\textsuperscript{1}, Ethan Steinberg\textsuperscript{1}, and Nigam H. Shah \textsuperscript{1,2,3,4}

\textsuperscript{1}Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, California, USA, \textsuperscript{2}Department of Medicine, School of Medicine, Stanford University, Stanford, California, USA, \textsuperscript{3}Clinical Excellence Research Center, Department of Medicine, Stanford University, Stanford, California, USA and \textsuperscript{4}Technology and Digital Solutions, Stanford Healthcare, Stanford, California, USA

Corresponding Author: Yizhe Xu, PhD, Stanford Center for Biomedical Informatics Research, Stanford University, 3180 Porter Dr., Palo Alto, CA 94304, USA; yizhex@stanford.edu

Received 11 October 2022; Revised 17 January 2023; Editorial Decision 21 January 2023; Accepted 11 February 2023

ABSTRACT

Objective: There are over 363 customized risk models of the American College of Cardiology and the American Heart Association (ACC/AHA) pooled cohort equations (PCE) in the literature, but their gains in clinical utility are rarely evaluated. We build new risk models for patients with specific comorbidities and geographic locations and evaluate whether performance improvements translate to gains in clinical utility.

Materials and Methods: We retrain a baseline PCE using the ACC/AHA PCE variables and revise it to incorporate subject-level information of geographic location and 2 comorbidity conditions. We apply fixed effects, random effects, and extreme gradient boosting (XGB) models to handle the correlation and heterogeneity induced by locations. Models are trained using 2,464,522 claims records from Optum\textsuperscript{\textregistered}s Clinformatics\textsuperscript{\textregistered} Data Mart and validated in the hold-out set (N = 1,056,224). We evaluate models' performance overall and across subgroups defined by the presence or absence of chronic kidney disease (CKD) or rheumatoid arthritis (RA) and geographic locations. We evaluate models' expected utility using net benefit and models' statistical properties using several discrimination and calibration metrics.

Results: The revised fixed effects and XGB models yielded improved discrimination, compared to baseline PCE, overall and in all comorbidity subgroups. XGB improved calibration for the subgroups with CKD or RA. However, the gains in net benefit are negligible, especially under low exchange rates.

Conclusions: Common approaches to revising risk calculators incorporating extra information or applying flexible models may enhance statistical performance; however, such improvement does not necessarily translate to higher clinical utility. Thus, we recommend future works to quantify the consequences of using risk calculators to guide clinical decisions.

Key words: atherosclerotic cardiovascular disease, clinical utility, model calibration, net benefit, pooled cohort equations, subgroup performance
INTRODUCTION
Atherosclerotic cardiovascular disease (ASCVD) and stroke are leading causes of death in the United States and cost healthcare systems about $330 billion each year. To manage these conditions, clinicians often evaluate patients’ risk of having a major adverse cardiovascular event using a risk estimator, to select a treatment strategy. The 2013 American College of Cardiology and the American Heart Association (ACC/AHA) pooled cohort equations (PCE) is the most commonly used ACSVD risk estimator, with the 2019 ACC/AHA ASCVD prevention guidelines specifying risk-based criteria to inform statin initiation. However, several studies have found that PCE is miscalibrated, which leads to substantial misestimation of patients’ ASCVD risk and further exacerbates health inequalities. In addition, Rodriguez and Khera pointed out that PCE overestimates the ASCVD risk for non-Hispanic whites, African Americans, Asians, Hispanics, and subjects with overweight and obesity, respectively.

Given the impact of the PCE’s specific risk estimates on ASCVD treatment decisions, a large number of efforts have been made to address this miscalibration by proposing revised models that utilize larger data sets or apply advanced machine learning approaches or fairness methods. Such efforts to revise the PCE may very well conduct to summarize the common risk scores (eg, FRS, QRISK, and SCORE) and modeling approaches, predictive models in particular populations, and how social determinants of health are used to help digest the large body of ASCVD risk estimation tools. It is now time to pause and ask the question that whether the revised risk models lead to changes in clinical outcomes or resource allocation.

Typically, the discrimination and calibration performance of risk prediction models are evaluated to summarize how well the individuals are ranked according to their risk, and how closely the predicted risks match the observed outcome rates, respectively. Although these 2 metrics are critical for assessing model reliability, they are agnostic of the way the model will be used for making decisions—as such, they do not directly correspond to the potential gains from using a risk model to guide an intervention according to prespecified criteria. Therefore, metrics that consider clinical consequences, for example, benefits and harms of an intervention, should be emphasized. For instance, the net benefit from decision-curve analysis and the incremental cost-effectiveness ratio from cost-effectiveness analysis are commonly used mechanisms to quantify the consequences of model updates on clinical decisions.

In this work, our primary goal is to evaluate whether the improved statistical performance shown in revised calculators also lead to higher expected clinical utility. We proceed such an evaluation through conducting comprehensive experiments in 2 steps: In the first step, we follow the standard process of creating revised calculators and first derive a baseline PCE by retraining the ACC/AHA PCE using the large claims data from Optum®’s Clininformatics Data Mart (CDM) ZIP5. We then revise the baseline PCE by including 3 additional patient-level features, including residence zip codes and 2 comorbidities, which results in a fixed effect model for predicting the 5-year ASCVD risk. In addition, we apply 2 alternative modeling approaches, a random effects model and an extreme gradient boosting (XGB) model. In the second step, we compare the performance of the baseline PCE and 3 revised models in terms of discrimination, calibration, and expected utility overall and in subgroups of interest. We adopt several metrics to measure model calibration at various granularity levels and measure clinical utility as net benefits using decision curve analysis for multiple exchange rates.

MATERIALS AND METHODS

Data sources and formulation of the study cohort
Optum CDM is a deidentified database derived from a large adjudicated claims data warehouse that contains over 51 million patients. This large database contains data on medical and pharmacy claims and lab results from 2003 to 2021. To evaluate the performance of adding geographic location information to the PCE equation, we utilize a variant of Optum data, the Optum ZIP database version 5.0, that includes 5-digit zip code information.

We apply similar eligible criteria described in the current clinical practice guidelines for statin initiation to extract the primary study cohort from Optum ZIP. First, to obtain an analysis data set that captures the ACC/AHA PCE variables, we remove all patients without high-density lipoprotein (HDL) or cholesterol measurements. In addition, we restrict the cohorts to patients between 40 and 75 years old with no prior history of cardiovascular disease events or statin use at the time of visit. We also require patients to have HDL between 20 and 100 mg/dL and total cholesterol between 130 and 320 mg/dL. We attempt to require systolic blood pressure (SBP) measurements, but are unable to find a sufficient number of patients with such lab result; so we instead replace all SBP variables with binary indications of prior hypertension diagnoses. Also, Optum ZIP does not contain race information. We refer to the final data set as the PCE-eligible cohort.

The Stanford medicine research data repository (STARR) is a clinical warehouse that contains approximately 130 million encounter records from the Stanford Health Care and Lucile Packard Children’s Hospital. This database comprises longitudinal data on patient and provider information, diagnoses, procedures, medication orders, and lab results that have been normalized to the Observational Medical Outcomes Partnership (OMOP) CDM version 5.3.1, sourced from inpatient and outpatient clinical encounters that occurred between 1990 and 2021. The use of this data is conducted in accordance with all relevant guidelines and regulations. Approval for the use of STARR for this study is granted by the Stanford Institutional Review Board Administrative Panel on Human Subjects in Medical Research (eProtocol #IRB-46829), with a waiver of informed consent.

In our study, the STARR data is used for 2 purposes: (1) To quantify the combined impact of using the hypertension indicator to replace the SBP variable and not using race as a predictor in the baseline PCE on its performance; (2) To serve as an external validation data set for the ACC/AHA PCE and retrained baseline PCE. To create the analysis data set, we follow the same algorithms used for Optum ZIP but include the continuous SBP variable and race.

For both Optum ZIP and STARR, we require every subject to have at least 1 year of data prior to their index date, and we generate the final cohort by randomly sampling one record for each patient from the resulting candidate index events. The sample sizes of the PCE-eligible cohorts from Optum ZIP and STARR are 3,520,746 and 24,440, respectively. The distribution of baseline covariates is summarized in Table 1.

Outcome and feature extraction
We study the composite ASCVD outcome by estimating the 5-year risk of a fatal coronary heart disease, myocardial infarction, or stroke event.
We consider coronary heart disease to be fatal if death occurs within a year after being diagnosed. The censoring occurs when a subject started using statin, died, or reached the end of the follow-up period (ie, when subjects switch to a difference insurance for Optum ZIP; the last record in the Stanford health system for STARR). For both data sources above, we create the ASCVD event indicator according to these definitions and compute the follow-up time as the time from an index date to the date of having an event or being censored.

We use OMOP CDM concept identifiers to extract time-agnostic demographic features (race and ethnicity, sex, and age), drug exposures (antihypertensives), lab test results (HDL, total cholesterol, SBP), and condition occurrences (diabetes mellitus, chronic kidney disease [CKD], rheumatoid arthritis [RA], hypertension, and smoking history), and the 5-digit zip code. After searching using Logical Observation Identifiers Names and Codes (LOINC), we find little SBP information in Optum ZIP due to the fact that Optum started collecting data on SBP around 2019. Thus, we extract the hypertension condition as a surrogate variable for SBP. In general, the lab results are not well recorded in claims and electronic health records databases, so we focus on the PCE-eligible cohort defined above instead of conducting multiple imputations as the assumption of missing at random may not hold in our situation. Important concepts and corresponding codes are presented in Supplementary Table S3 using the Systematized Nomenclature of Medicine (SNOMED) concept codes.

Model learning
For all the model construction below, we randomly sample (with the same seed) 70% of the Optum ZIP data as the training set and use the rest as the testing set.

The baseline PCE
The 2013 ACC/AHA PCE were derived using Cox proportional hazard (PH) models for 4 demographic groups defined by sex (female vs male) and race and ethnicity (African American vs White). Due to the limited data in Optum ZIP, we retrain the ACC/AHA PCE using a similar set of covariates where SBP is replaced by an indicator of prior hypertension condition and the race variable is left out. Vyas et al$^{28}$ also pointed out that using race for risk estimation may worsen racial inequity. Finally, we fit 1 Cox PH model with sex being one of the covariates instead of developing 2 separate models for male and female groups as in the ACC/AHA PCE, that is,

$$\lambda(t) = \hat{\lambda}(t) \exp(\beta^T X_i),$$

where $\hat{\lambda}(t)$ and $\lambda(t)$ are the baseline hazard and the hazard of subject $j$ at time $t$, respectively, and $\beta$ is the vector of coefficients (effects on the log hazard scale) of the PCE variables. The 5-year ASCVD risk is then computed as $1 - S(t)$, where $S(t) = \exp(-H(t))$ with the cumulative hazard $H(t) = \int_0^t \hat{\lambda}(u) du$ and $t = 5$ years. This retrained model is used as the surrogate of the ACC/AHA PCE and referred to as baseline PCE. Furthermore, we perform a sensitivity analysis using the STARR data (external validation) to quantify the difference between baseline PCE and ACC/AHA PCE.

The revised PCE
Given the consistent prior evidence on the ACC/AHA PCE being inaccurate, we revise the baseline PCE by utilizing additional subject-level information on comorbidities and geographic locations. CKD and RA are 2 risk-enhancing factors$^{29,30}$ that are recommended to consider for adults at borderline (5–7.5%) and intermediate risk (7.5–20%),$^3$ so we include them as binary variables in multivariate models.

Individuals who reside in the same area may share common environmental (eg, air pollution), lifestyle (eg, diet and physical activity), and other social determinants of heath factors, so their survival outcomes may be positively dependent, that is, the random errors that are absorbed into the baseline hazard $\lambda_0(t)$ are not independent and identically distributed (i.i.d.). In addition, the differences across local areas may contribute to the unobserved heterogeneity, which may improve individual-level risk estimation when taken into consideration. Thus, we also incorporate individuals’ zip code information to enable local-area risk estimation. To deal with the issue that some location subgroups have small populations, we group subjects using the code of sectional center facility (ie, the first 3 digits of the 5-digit zip code) and merge a small group with the nearest location iteratively to ensure a minimum sample size of 3000 within each location subgroup. The distribution of merged location is displayed in Supplementary Figure S2. To account for the correlation and heterogeneity issues induced by locations, we construct the following 3 revised models: First, a fixed effects model that includes the location variable as a factor, that is,

$$\lambda(t) = \hat{\lambda}(t) \exp(\gamma_j^T X_i),$$

where $i$ and $j$ are the location and individual indices, respectively, and $\gamma_j$ denotes the average (constant) effects for location $i$. $X_i$ represents the baseline PCE covariates of subject $j$ in the location $i$ and $\beta$ is the vector of corresponding effect estimates. With a Cox regression, the first location is used as the reference level, that is, $\gamma_1 = 0$, and the effects in other locations compared to the first location are identical. In addition, the differences across local areas may contribute to the unobserved heterogeneity, which may improve individual-level risk estimation when taken into consideration. Thus, we also incorporate individuals’ zip code information to enable local-area risk estimation. To deal with the issue that some location subgroups have small populations, we group subjects using the code of sectional center facility (ie, the first 3 digits of the 5-digit zip code) and merge a small group with the nearest location iteratively to ensure a minimum sample size of 3000 within each location subgroup. The distribution of merged location is displayed in Supplementary Figure S2. To account for the correlation and heterogeneity issues induced by locations, we construct the following 3 revised models: First, a fixed effects model that includes the location variable as a factor, that is,

$$\lambda(t) = \hat{\lambda}(t) \exp(\gamma_j^T X_i),$$

where $i$ and $j$ are the location and individual indices, respectively, and $\gamma_j$ denotes the average (constant) effects for location $i$. $X_i$ represents the baseline PCE covariates of subject $j$ in the location $i$ and $\beta$ is the vector of corresponding effect estimates. With a Cox regression, the first location is used as the reference level, that is, $\gamma_1 = 0$, and the effects in other locations compared to the first location are the coefficients $\gamma_j$, $i \neq 1$, by the model.

Second, a random effects model by assuming the location subgroups are random draws from a larger population of locations, that is,

$$\lambda(t) = Z_i \lambda(t) \exp(\beta^T X_i),$$

where $Z_i$ is the random location effect that follows a gamma distribution in our case. In this model, we assume $Z$ is shared among all

---

**Table 1. Summary of baseline characteristics in the Optum ZIP and STARR PCE-eligible cohorts**

| Variable                  | Optum ZIP (N = 3 520 746) | STARR (N = 24 440) |
|---------------------------|---------------------------|--------------------|
| Age (years)               | 53.5 (9.9)                | 55.7 (9.4)         |
| Female                    | 2 116 013 (57.4%)         | 14 084 (57.6%)     |
| Current smoker            | 328 486 (8.9%)            | 1489 (6.1%)        |
| High-density lipoprotein  | 55.8 (17.4)               | 58.5 (16.4)        |
| Total cholesterol         | 197.0 (3.4)               | 195.2 (3.1)        |
| Antihypertensives         | 1 081 759 (29.3%)         | 4005 (16.4%)       |
| Diabetes                  | 499 243 (13.5%)           | 1696 (6.9%)        |
| Hypertension              | 1 439 459 (39.0%)         | 6514 (26.7%)       |
| Chronic kidney disease    | 109 116 (3.0%)            | 580 (2.4%)         |
| Rheumatoid arthritis      | 64 641 (1.8%)             | 225 (0.9%)         |
| Systolic blood pressure   | —                         | 126.1 (16.6)       |
| Race and ethnicity        | —                         | 22 511 (92.1%)     |
| Black or African American | —                         | 1929 (7.9%)        |
| White                     | —                         | 22 511 (92.1%)     |

*Note: Continuous variables are summarized as mean and standard deviation and categorical variables are summarized as count and percentage. The total sample sizes (including ineligible patients) of Optum ZIP and STARR are 16 792 476 and 24 440, respectively.*
the subjects within the same location \( i \), and it is constant over time. Compared to the fixed effects model, the random effects model shrinks the local estimates towards the mean estimates over all locations, which helps to improve inferences when the sample sizes of some location subgroups are small.

Third, an XGB model for estimating nonlinear relationships between covariates and outcome from a data-driven perspective.\(^{31}\) XGB is a scalable and fast implementation of the gradient boosting machine method that handles sparse data and instance weights in approximate tree learning. The detailed explanation of the XGB approach is given in Supplementary Appendix S1.1. To obtain a top approximate tree learning, the fractions of row and column subsampling, and the best set of hyperparameters (1000, 3, 0.9, 1, respectively) is chosen through grid search using 5-fold cross-validation. As a slower learning rate often leads to better estimation performance but requires a large number of iterations, we use a learning rate of 0.1 with 100 trees (a.k.a iterations) during the hyperparameter tuning process and reduce it to 0.05 with 500 trees in the final model. The optimal number of boosting iterations is determined by applying early stopping, that is, the model training is stopped if the negative partial log-likelihood in the validation set is not improved for 20 rounds. Our analyses using XGB are carried out by the R package \texttt{xgboost}.

**Evaluation metrics for risk models**

We evaluate the overall and subgroup performance of the risk models in “Model learning” section using 3 metrics: discrimination, calibration, and net benefit. As the first 2 metrics are standard, we include the details in Supplementary Appendix S1.2.

**Clinical utility metrics**

To measure the clinical impact of using a risk model to guide interventions, we conduct decision curve analysis and compare model performance using net benefit that considers the trade-off between benefits and harms of an intervention.\(^{28,35}\) Net benefit quantifies the clinical utility of a model by incorporating the clinical consequence, specified by an exchange rate, into the calculation. For instance, the exchange rate of 20% corresponds to an accepted tradeoff to treat 5 subjects with statin in order to prevent one ASCVD event. This tradeoff is determined by cost-benefit analysis, and corresponds to a risk level where expected benefits of a given statin regime outweigh the harms of side effects.\(^{3}\) So, the exchange rate can be viewed as a threshold of ASCVD risk used to determine whether or not to initiate statin. Specifically, the net benefit (NB) are computed as

\[
NB = \left( \frac{TP}{N} - \frac{FP}{N} \right) \times \frac{p_t}{1 - p_t},
\]

where \( \lfloor . \rfloor \) denotes the number of observations that satisfy a condition, and TP and FP are short for true and false positives, respectively. The probability threshold is denoted as \( p_t \), and a subject is treated if the predicted risk \( R > p_t \) and untreated otherwise. A decision curve is drawn by plotting the net benefit for a sequence of unique thresholds, and we focus on three 5-year ASCVD risk thresholds of 0.25%, 0.375%, and 10%, which is equivalently to the 0.5%, 0.75%, and 20% 10-year probability thresholds recommended by the ASCVD prevention guideline.\(^{3}\)

The net benefit is in the unit of true positives, so a value of 5% indicates that a risk prediction model leads to correctly treating 5 patients per 100 at risk without increasing the number of false interventions. When comparing different risk models, we may also convert the difference in net benefit to the number of additional true positives or false positives that one model can yield or prevent compared to the other, respectively. Let \( a = \lfloor TP/N \rfloor, b = \lfloor FP/N \rfloor, \) and \( k = p_t/(1 - p_t) \), then given a threshold, suppose the net benefit for the baseline PCE is \( NB_0 = a - bk \) and for a revised model is \( NB = a - bk \), then we have

\[
NB - NB_0 = (a - a_0) + (b_0 - b)k.
\]

Thus, if the revised model achieves the same false positive rate as the baseline PCE, that is, \( b = b_0 \), it correctly treats additional 1000(NB – NB\(_0\)) patients per 1000 subjects at risk. Likewise, if the true positive rates are the same, that is, \( a = a_0 \), the revised model prevents \( b_0 - b = 1000(NB - NB_0)/k \) unnecessary treatment per 1000 subjects at risk.

**RESULTS**

To make it easier to understand the differences between models, we report the difference in performance between each of our revised risk models and the baseline PCE. We also compute 95% confidence intervals on these differences using bootstrapping where the difference in performance (eg, a C statistic from a revised model minus that of the baseline PCE) is calculated within each bootstrap sample. We present the results from internal validation in the Optum testing data below and show the external validation results in the STARR data in Supplementary Appendix S2.2.

**Overall performance**

Table 2 shows how the overall performance of each revised model compares to the performance of the baseline PCE. All revised models resulted in higher C statistics than baseline PCE, with the fixed effects model yielding the largest improvement. We evaluated the calibration-in-the-large performance by comparing bias calculated as the absolute differences between estimated O/E and the ideal value of 1, and we see all models performed similarly. According to the GND test, the baseline PCE and random effects models were miscalibrated \( (P < .05) \), but the revised fixed effects and XGB models showed insignificant \( P \)-values, indicating no evidence of miscalibration (Supplementary Table S1). In addition, the pairwise tests showed that the XGB model yielded a smaller GND test statistic, which indicates an improvement in model calibration. The calibration plots in Figure 1 also showed that the estimates from the XGB model were closer to the ideal reference line (dotted line) than those from the baseline PCE. We also compared the biases in calibration slopes (absolute differences between estimates and 1) and found no significant improvement.

In terms of expected clinical utility, the differences in net benefit between the revised models and baseline PCE were small. If a doctor is willing to prescribe statin to 40 subjects in order to correctly treat one patient who will develop ASCVD within 5 years (threshold of 2.5%), the net benefit of baseline PCE and XGB was 0.0736 (Supplementary Table S1), which means that these 2 risk estimators are equivalent to a strategy that correctly treats 74 patients per 1000 at risk without increasing the number of false interventions. When the thresholds are 3.75% and 10%, the fixed effects model resulted in slightly higher net benefit than the baseline PCE (difference = 0.0004), which maps to correctly treating 4 more patients.
than the baseline PCE per 10000 patients at risk if the false positive rates are the same, or preventing 103 and 36 unnecessary treatment per 10000 at risk, respectively, if the true positive rates are identical (Table 2). Supplementary Figure S3 showed the decision curves for all risk estimators that capture the net benefit under a range of risk thresholds from 0% to 10%. We see that the 4 risk estimators provided nearly identical net benefit, but all of them achieved higher utility than treating everyone in the population when the probability

### Table 2. Difference in overall performance between the revised models and the baseline PCE when estimating 5-year ASCVD risk

| Evaluation | Fixed effects | Random effects | XGB |
|------------|---------------|----------------|-----|
| C statistic | 0.0096 (0.0088, 0.0104) * | 0.0003 (0.0001, 0.0005) | 0.0033 (0.0028, 0.0038) |
| O/E        | 0.0003 (−0.0006, 0.0010) | 0.0002 (−0.0002, 0.0004) | 0.0004 (−0.0003, 0.0009) |
| GND        | −21.6 (−46.4, 3.9) | 60.4 (33.1, 87.7) | −21.7 (−44.3, −0.7) * |
| Calibration slope | 0.0020 (−0.0105, 0.0118) | 0.0218 (−0.0105, 0.0313) | 0.0015 (−0.0125, 0.0130) |
| NB (2.5%)  | −0.0003 (−0.0004, −0.0002) | −0.0001 (−0.0002, −0.0001) | 0.0001 (0.0000, 0.0001) * |
| NB (3.75%) | 0.0004 (0.0002, 0.0006) * | 0.0001 (0.0000, 0.0002) | 0.0002 (0.0001, 0.0004) |
| NB (10%)   | 0.0004 (0.0000, 0.0007) * | −0.0001 (−0.0003, 0.0001) | 0.0002 (0.0000, 0.0005) |

NB: net benefit.

*Indicates that a particular method both had the best performance and that the performance improvement over the baseline PCE was statistically significant. The XGB model improved both discrimination and calibration performance but only slightly increased the net benefit.

[Figure 1. Overall calibration plots of baseline PCE and revised models. Observed and estimated risks are compared using 135 risk bins. The fixed effects model and XGB model outperformed the baseline PCE.]
CKD is observed in the presence of RA subgroup under the risk threshold similarly.

NB (2.5%) 0.0003 (0.0004, −0.0001) 0.0001 (−0.0002, 0.0001) 0.0001 (0.0000, 0.0001)* 0.0001 (0.0000, 0.0001)*

NB (3.75%) 0.0004 (0.0002, 0.0007)* 0.0001 (0.0000, 0.0002) 0.0002 (0.0000, 0.0004)

NB (10%) 0.0004 (0.0001, 0.0008)* −0.0001 (−0.0003, 0.0001) 0.0002 (−0.0001, 0.0004)

CKD = 1

C statistic 0.0076 (0.0019, 0.0131) −0.0012 (−0.0025, 0.0000) 0.0096 (0.0052, 0.0138)∗

O/E −0.0723 (−0.0998, −0.0015) −0.0485 (−0.1310, 0.0328) −0.0727 (−0.0933, −0.0668)∗

GND −13.5 (−32.3, 7.0) −6.0 (−19.4, 10.2) −25.7 (−43.8, −7.1)∗

Calibration slope 0.0075 (0.0459, 0.0646) 0.0530 (0.3312, 0.0742) −0.0604 (−0.1122, −0.0666)∗

NB (2.5%) −0.0002 (−0.0005, 0.0000) 0.0001 (0.0000, 0.0001)* 0.0001 (0.0000, 0.0001)*

NB (3.75%) −0.0002 (−0.0007, 0.0003) 0.0001 (−0.0002, 0.0005) 0.0003 (−0.0001, 0.0007)

NB (10%) 0.0001 (−0.0020, 0.0022) 0.0001 (−0.0011, 0.0015) 0.0027 (0.0005, 0.0052)*

RA = 0

C statistic 0.0092 (0.0083, 0.0102)* 0.0000 (−0.0001, 0.0001) 0.0029 (0.0024, 0.0034)

O/E 0.0046 (−0.0065, 0.0076) 0.0039 (−0.0061, 0.0063) 0.0045 (−0.0066, 0.0072)

GND −17.9 (−45.5, 7.4) 56.8 (25.0, 86.9) −18.5 (−43.6, 3.3)

Calibration slope 0.0037 (−0.0162, 0.0192) 0.0272 (0.0080, 0.0379) 0.0003 (−0.0075, 0.0080)

NB (2.5%) −0.0003 (−0.0004, −0.0002) −0.0001 (−0.0002, −0.0001) 0.0001 (0.0000, 0.0001)*

NB (3.75%) 0.0004 (0.0002, 0.0006)* 0.0001 (0.0000, 0.0002) 0.0002 (0.0000, 0.0003)

NB (10%) 0.0003 (0.0000, 0.0007) −0.0001 (−0.0003, 0.0001) 0.0002 (0.0000, 0.0005)

RA = 1

C statistic 0.0137 (0.0075, 0.0200)* −0.0006 (−0.0016, 0.0005) 0.0051 (0.0019, 0.0084)

O/E −0.2689 (−0.3109, −0.1635)* −0.2622 (−0.2836, −0.1873) −0.2688 (−0.3030, −0.1707)

GND −37.9 (−67.5, −5.7) −23.4 (−40.7, −3.0) −39.1 (−63.9, −13.6)*

Calibration slope 0.0931 (0.0281, 0.1564) 0.1336 (0.0928, 0.1621) 0.0964 (0.0420, 0.1428)

NB (2.5%) 0.0000 (−0.0001, 0.0002) 0.0000 (−0.0001, 0.0001) 0.0000 (−0.0001, 0.0001)

NB (3.75%) 0.0010 (−0.0006, 0.0028) 0.0010 (−0.0004, 0.0025) 0.0019 (0.0000, 0.0040)*

NB (10%) 0.0040 (0.0000, 0.0082)* 0.0012 (−0.0015, 0.0043) 0.0024 (−0.0013, 0.0063)*

Note: Subgroup CKD = 0 includes patients with the absence of CKD condition regardless of the status of RA condition, and the other subgroups are defined similarly. NB (2.5%) indicates the net benefit is evaluated using a threshold of 2.5%, and others can be interpreted similarly.

*Indicates that a particular method both had the best performance and that the performance improvement over the baseline PCE was statistically significant.

The revised fixed effects and XGB models showed improved discrimination in all subgroups, and XGB yielded better calibration in the subgroups of presence of CKD or RA. However, the gains in net benefit are tiny in general.

Threshold is above 4%. The overall performance on the original scales is shown in Supplementary Table S1.

Subgroup performance

By comorbidity

Table 3 shows the performance across subgroups defined by the absence or presence of CKD and RA conditions. The fixed effects and XGB models resulted in higher C statistics than the baseline PCE in all subgroups. Furthermore, the revised models yielded less biased O/E ratios and smaller GND statistics than the baseline PCE in the small subgroups defined by the presence of CKD or RA, which indicates that there was an improvement in model calibration. The revised models yielded similar calibration slope estimates as baseline PCE. The subgroup performance on the original scales is shown in Supplementary Table S2.

For expected clinical utility, Figure 2 shows that the net benefit in the subgroups with a certain comorbidity was much higher than that in the subgroups without that comorbidity due to higher ASCVD incidence rates. Although the revised XGB and fixed effects models yielded larger net benefit than baseline PCE in some subgroups, the differences are tiny (Table 3). The largest improvement is observed in the presence of RA subgroup under the risk threshold of 10% (difference = 0.0040), which indicates that the fixed effects model may correctly treat 4 more patients than the baseline PCE per 1000 at risk if the false positive rates are assumed to be the same, or prevent 36 unnecessary treatment per 1000 at risk if the true positive rates are the same.

By geographic location

Given the difficulty in tabulating the performance of 299 locations, in Figure 3, we summarize the statistical performance using density plots of the difference from baseline for 3 metrics, the C statistic, O/E values, and calibration slope. The C statistic panel shows that the XGB model outperformed baseline PCE in most of the locations although the magnitudes of improvements were small. The O/E panel suggests that the random effects and XGB models performed similarly as baseline PCE for most of the locations; however, the fixed effects model outperformed the baseline by showing various amounts of bias reduction in O/E estimates for many locations. Finally, the revised models yielded similar calibration slopes as the baseline PCE across all locations.

Figure 4 displays the difference in net benefit of revised models as compared to the baseline PCE across different local areas. The gains in net benefit were nearly zero when the risk threshold was
0.025. As the threshold increased, both the gains and losses from revised models increased, but the distributions were still centered around zero indicating that no major utility gains in general. Supplementary Figure S1 shows the differences in net benefit between the fixed effects model and baseline PCE in locations that contain fewer than or equal to 1000 subjects. We see positive utility gains for 9 out of 14 locations at both the thresholds of 0.0375 and 0.1, suggesting improved net benefit for areas with small populations. However, this result may not generalize given our small sample size (14 locations), the small gains we observed, and that we only considered a single cutoff of 1000.

**DISCUSSION**

Using a large insurance claims dataset, OptumC’s CDM ZIP5, that contains over 3.5 million patient records collected from all 50 states.
in the United States, we conducted a comprehensive evaluation of several ASCVD risk models to examine whether improved statistical performance maps to gains in clinical utility. Compared to the baseline PCE, our revised fixed effects and XGB models resulted in improved discrimination overall and in all comorbidity subgroups, and XGB enhanced model calibration in the underrepresented subgroups with presence of CKD or RA. Additionally, the fixed effects model and XGB model yielded better calibration-in-the-large and discrimination, respectively, in most of the geographical locations. However, we found that even in subgroups where improved calibration or discrimination was observed, the net benefit yielded was small in general. Our experiments illustrate that improvement in terms of statistical performance metrics does not necessarily translate to meaningfully higher expected utility.

Our results suggest that, for some treatment thresholds, it may not be necessary to revise the baseline PCE, if the goal is to maximize the expected clinical utility for a general population. However, in particular cases, revising the baseline PCE by using additional patient-level information and applying flexible statistical models may result in larger net benefit (eg, for the high-risk subgroups with the presence of CKD or RA and when the 5-year ASCVD risk threshold is above 10%). In addition, although we found that model revision may help improve the expected utility in subgroups with small populations, the definition of “small-sized local areas” may not generalize to every context since we only had 14 such locations and the magnitudes of gains were small. This result underscores the core argument that it is necessary to always assess whether clinical gains are also observed in situations where revised models showed improved statistical performance. Also, future works are needed to further investigate the expected utility in geographical areas with varied sizes.

Given the large amount of prior work where conclusions are drawn primarily based on statistical performance, we highlight that net benefit is a unique, direct measure of the consequences from informing clinical decisions using risk stratification tools. Net benefit allows flexibly weighing the gain from correctly treating patients who will develop the disease, against the loss from assigning unnecessary treatment to subjects who will not develop the disease, in light of stakeholders’ preferences (or cost-benefit analysis that quantitatively compares harms and benefits of treatment\(^{32-34}\)). Thus, showing such a mismatch between statistical performance and clinical utility is important for encouraging future work to rethink how to define a “useful,” or improved, model. To facilitate interpretation, we converted the values of net benefit into the additional number of patients being correctly treated or the number of unnecessary interventions being prevented per 1000 subjects at risk. This interpretation is especially useful when comparing multiple risk stratification tools.

Prior theoretical results found that clinical utility depends on model’s calibration\(^ {7,35,36}\) and better calibrated models often have larger clinical benefits.\(^ {37}\) Our results from the comorbidity subgroup evaluations agree with prior findings and showed larger gains in net

---

**Figure 4.** Net benefit of revised models compared to the baseline PCE by geographic locations. FE and RE are short for fixed effects and random effects models, respectively. The revised fixed effects and XGB models yielded similar net benefit across all 3 risk thresholds.
benefit when model calibration is significantly improved. In addition, we noticed that a model with degraded calibration also has decreased net benefit even if the discrimination has been improved. This implies that net benefit is associated more closely with calibration than discrimination performance.

We also observed large heterogeneity in model calibration across subgroups, which is referred to as metric disparity in model fairness assessment. As expected, such metric disparity is exacerbated when the number of subgroups grows larger, such as across 299 geographic locations. Prior works attempted to diminish such fairness violation by incorporating fairness constraints into a training objective function but resulted in worse statistical performance and reduced net benefit in general.

In the case of ASCVD risk estimation, Foryciarz et al. showed that fairness evaluations should focus on comparisons of local calibration at thresholds rather than on classification errors, because of the assumptions on harm and benefit underlying the treatment rules (where benefit of statin treatment scales with the underlying risk, by bringing down low-density lipoprotein cholesterol levels). This suggests that further incorporating metrics of clinical consequence that directly consider expected risk and benefit, rather than implicitly through changes in clinical decisions, might be valuable for comparing model fairness.

Our evaluation of the ASCVD risk models has the following strengths: First, we utilized the real-world and large claims data, Optum’s CDM ZIP5, that contains subject-level information on comorbidities and residence zip code. The latter enabled us to evaluate model performance across a large number of geographic locations with some confidence in the sense that each location subgroup contains at least 900 patients in the testing data. Second, we focused on comparing models’ clinical utility, so our results directly inform the gains obtained from deploying revised models to guide interventions. In addition, we evaluate net benefit under multiple thresholds to reflect various decision-makers’ perceptions of the benefits of intervening on a true case over the harms of unneeded treatment. Third, we additionally evaluated our risk models in terms of statistical performance to illustrate the relationship between the clinical utility metrics and more standard performance metrics in our experimental setting.

Our study has several limitations. First, SBP and race information is missing in the Optum ZIP data. This weakness forced us to use a surrogate baseline that underperformed the 2013 ACC/AHA PCE. Despite this limitation, our comparison of revised models to the baseline PCE is valid, since all models were trained using the Optum ZIP data with the same data restriction. Second, we combined some of the original locations defined by 5-digits zip code, to avoid subgroups with extremely small sizes. Even though the characteristics within each merged location may become more heterogeneous, merging small areas was necessary to ensure sufficient sample sizes that are critical for conducting reliable subgroup evaluations. As we combined each location with its nearest neighbours iteratively until the minimum sample size was reached, the correlation within each original local area was maximally preserved. Third, due to computation burden, we used a relatively simple discrimination metric that assumes noninformative censoring. However, our experiments using the IPCW C statistic, which allows modeling censoring mechanism as a function of covariate adjustment, yielded very similar results.

Although we emphasized the ASCVD risk stratification tools in this work, our model evaluation framework focused on clinical utility can be easily applied to many other risk calculators such as CHA2DS2VASc score for atrial fibrillation stroke and the Model for End-Stage Liver Disease (MELD) score. We appreciate the large amount of prior efforts on refining clinical risk estimators (eg, the ones included in MDcalc), but we believe that future work needs to examine model improvements in terms of better clinical consequences. Our study can be viewed as such an example where we determine whether or not a revised model is worth being deployed based on its clinical value, that is, its consequences on guiding interventions.

CONCLUSION
Revising clinical risk estimators using additional information or advanced statistical models may help to improve statistical performance. However, such improvements do not necessarily translate into meaningful gains in clinical utility. Thus, we recommend that future efforts on improving risk stratification tools should include evaluation metrics that consider clinical consequences.

FUNDING
This work was supported by R01 HL144555 from the National Heart, Lung, and Blood Institute (NHLBI).

AUTHOR CONTRIBUTIONS
YX: data extraction, analysis design, analysis implementation, and drafting the manuscript. AF: data extraction and manuscript review. ES: data extraction and manuscript review. NHS: supervising and manuscript review.

SUPPLEMENTARY MATERIAL
Supplementary material is available at Journal of the American Medical Informatics Association online.

ACKNOWLEDGMENTS
We appreciate the great help from Dr Stephen Pfohl on data extraction and reviewing manuscript.

CONFLICT OF INTEREST STATEMENT
None declared.

DATA AVAILABILITY
Data for this project were accessed using the Stanford Center for Population Health Sciences Data Core. The PHS Data Core is supported by a National Institutes of Health National Center for Advancing Translational Science Clinical and Translational Science Award (UL1TR003142) and from Internal Stanford funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES
1. Branchard B, Deb-Rinker P, Dubois A, et al. At-a-glance-how healthy are Canadians? A brief update. Health Promot Chronic Dis Prev Can 2018;38 (10): 385–7.
2. Andrus B, Lacaille D. ACC/AHA guideline on the assessment of cardiovascular risk. J Am Coll Cardiol 2013; 63 (25 Pt A): 2886.
3. Arnett DK, Blumenthal RS, Albert MA, et al. ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. *J Am Coll Cardiol* 2019; 74 (10): e177–232.

4. Cook NR, Ridker PM. Calibration of the pooled cohort equations for atherosclerotic cardiovascular disease: an update. *Ann Intern Med* 2016; 165 (11): 786–94.

5. Yadilowsky S, Hayward RA, Sussman JB, McClelland RL, Min Y-I, Basu S. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann Intern Med* 2018; 169 (1): 20–9.

6. Damen JA, Pajouheshnia R, Heus P, et al. Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. *BMJ* 2019; 17 (1): 1–16.

7. Foryciarz A, Pfohl SR, Patel B, Shah N. Atherosclerotic cardiovascular disease risk prediction in disaggregated Asian and Hispanic subgroups using electronic health records. *J Am Heart Assoc* 2019; 8 (14): e011874.

8. Rodriguez F, Chung S, Blum MR, Coulet A, Basu S, Panaliapannap LN. Atherosclerotic cardiovascular disease risk prediction in disaggregated Asian and Hispanic subgroups using electronic health records. *J Am Heart Assoc* 2019; 8 (14): e011874.

9. Khera R, Pandey A, Ayers CR, et al. Performance of the pooled cohort equations to estimate atherosclerotic cardiovascular disease risk by body mass index. *JAMA Netw Open* 2020; 3 (10): e2032342.

10. Wallisch C, Heinze G, Rinner C, et al. Re-estimation improved the performance of two Framingham cardiovascular risk equations and the pooled cohort equations: a nationwide registry analysis. *Sci Rep* 2020; 10 (1): 1–11.

11. Barda N, Yona G, Rothblum GN, et al. Addressing bias in prediction models by improving subpopulation calibration. *J Am Med Inform Assoc* 2021; 28 (3): 549–58.

12. Damen JA, Hoft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016; 353: i2416.

13. Mohd Faizal AS, Malathi Thevarajah T, Mei Khor S, Chang S-W. A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. *Comput Methods Programs Biomed* 2021; 207: 106190.

14. Cai R, Wu X, Li C, Hao J. Prediction models for cardiovascular disease risk in the hypertensive population: a systematic review. *J Hypertens* 2020; 38 (9): 1632–9.

15. Zhao Y, Wood EP, Mirin N, Cook SH, Chunara R. Social determinants in machine learning cardiovascular disease prediction models: a systematic review. *Am J Prev Med* 2021; 61 (4): 596–605.

16. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15 (4): 361–87.

17. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; 23 (13): 2109–23.

18. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat Med* 2013; 32 (13): 2173–84.

19. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. *Stat Med* 2015; 34 (10): 1659–80.