Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation

Agata Foryciarz, Stephen R Pfohl, Birju Patel, Nigam Shah

ABSTRACT

Objectives The American College of Cardiology and the American Heart Association guidelines on primary prevention of atherosclerotic cardiovascular disease (ASCVD) recommend using 10-year ASCVD risk estimation models to initiate statin treatment. For guideline-concordant decision-making, risk estimates need to be calibrated. However, existing models are often miscalibrated for race, ethnicity and sex based subgroups. This study evaluates two algorithmic fairness approaches to adjust the risk estimators (group recalibration and equalised odds) for their compatibility with the assumptions underpinning the guidelines’ decision rules. Methods Using an updated pooled cohorts data set, we derive unconstrained, group-recalibrated and equalised odds-constrained versions of the 10-year ASCVD risk estimators, and compare their calibration at guideline-concordant decision thresholds.

Results We find that, compared with the unconstrained model, group-recalibration improves calibration at one of the relevant thresholds for each group, but exacerbates differences in false positive and false negative rates between groups. An equalised odds constraint, meant to equalise error rates across groups, does so by miscalibrating the model overall and at relevant decision thresholds.

Discussion Hence, because of induced miscalibration, decisions guided by risk estimators learned with an equalised odds fairness constraint are not concordant with existing guidelines. Conversely, recalibrating the model separately for each group can increase guideline compatibility, while increasing intergroup differences in error rates. As such, comparisons of error rates across groups can be misleading when guidelines recommend treating at fixed decision thresholds.

Conclusion The illustrated tradeoffs between satisfying a fairness criterion and retaining guideline compatibility underscore the need to evaluate models in the context of downstream interventions.

INTRODUCTION

While risk stratification models are central to personalising care, their use can worsen health inequities. In an effort to mitigate harms, several recent works propose algorithmic group fairness—mathematical criteria which require that certain statistical properties of a model’s predictions not differ across groups. However, identifying which statistical properties are most relevant to fairness in a given context is non-trivial. Hence, before applying fairness criteria for evaluation or model adjustment, it is crucial to examine how the model’s predictions will inform treatment decisions—and what effect those decisions will have on patients’ health.

Here, we consider the 2019 guidelines of the American College of Cardiology and the American Heart Association (ACC/AHA) on primary prevention of atherosclerotic cardiovascular disease (ASCVD), which codify...
the use of 10-year ASCVD risk predictions to inform a clinician-patient shared decision-making on initiating statin therapy. These guidelines recommend that individuals estimated to be at intermediate risk (>7.5%–20%) be considered for initiation for moderate- to high-intensity statin therapy, and that those at high risk (>20%) be considered for high-intensity statin therapy. Individuals at borderline risk (>5%–7.5%) may be considered for therapy under some circumstances. These therapeutic thresholds were established based on randomised control trials, and correspond to risk levels where expected overall benefits derived from low-density lipoprotein cholesterol reduction outweigh risks of side effects (online supplemental file C). In general, such thresholds can be identified using decision analysis methods (figure 1A). The models accompanying the guidelines (pooled cohort equations, PCEs), developed for Black women, white women, Black men and white men, differ in both calibration and discrimination across groups. The resultant systematic bias in risk misestimation in these subgroups can lead to inappropriate or misinformed treatment decisions. Since then, several works derived updated equations some explicitly incorporating fairness adjustments.

If therapeutic thresholds recommended by guidelines reflect a balance of relevant harms and benefits for all subgroups, therapeutic decisions could be unfair if thresholds used for different groups differ, as they would lead to suboptimal treatment decisions for some groups (figure 1A). As such, subgroup calibration at optimal therapeutic thresholds is an important fairness criterion for 10-year ASCVD risk estimation models, since under miscalibration (systematic overestimation or underestimation of risk), treatment thresholds implicitly change from treatment thresholds to implied thresholds. An alternative fairness criterion, known as equalised odds (EO), which has previously been used to evaluate several clinical predictive models, requires equality in false positive and false negative error rates (FPR and FNR) across groups. One work proposed to explicitly incorporate EO constraints into the training objective to learn ASCVD risk estimators with minimal intergroup differences. In the context of ASCVD risk estimation, the EO criterion lacks a clear motivation and can thus yield misleading results. FPR and FNR are sensitive to the distribution of risk and are expected to differ across groups when the incidence of outcomes differs (figure 1B).
Furthermore, approaches to build EO-satisfying models either explicitly adjust group-specific decision thresholds, introduce differential miscalibration or reduce model fit for each group—which may lead to suboptimal decisions (figure 1A,C). EO-satisfying models may therefore be less appropriate than calibrated estimators for use with the ACC/AHA guidelines.17

We aim to evaluate the tension between calibration, EO and guideline-concordant decision-making. To do so, we propose a measure of local calibration at guideline-concordant therapeutic thresholds as a method for probing guideline compatibility and apply it to unconstrained, group-recalibrated and EO-constrained versions of the 10-year ASCVD risk prediction models learnt from the updated pooled cohorts data set,9,11 as well as the original8 and revised PCEs.11 We assess the proposed local calibration measure and error rates across groups for each model, and conclude with recommendations for identifying quantification and adjustment criteria for enabling fair model-guided decisions.

METHODS
Data sets
We use an updated pooled cohorts data set,11 comprised of ARIC (Atherosclerosis Risk in Communities Study, 1987–2011), CARDIA (Coronary Artery Risk Development in Young Adults Study, 1983–2006), CHS (Cardiovascular Health Study, 1989–1999), FHS OS (Framingham Heart Study Offspring Cohort, 1971–2014), MESA (Multi-Ethnic Study of Atherosclerosis, 2000–2012) and JHS (Jackson Heart Study, 2000–2012). Following the original PCE inclusion criteria,9 we include individuals aged 40–79, excluding those with a history of myocardial infarction, stroke, coronary bypass surgery, angioplasty, congestive heart failure or atrial fibrillation, or receiving statins at the time of the initial examination. We include all individuals, regardless of racial category, and classify them as Black and non-Black, consistent with the use of PCEs in practice for non-Black patients of colour.

We extract features included in the PCEs (total cholesterol, high-density lipoprotein (HDL) cholesterol, treated and untreated systolic blood pressure, diabetes, smoking status, age, binary sex and race) and body mass index, recorded at the initial examination. We also extract dates of observed ASCVD events (myocardial infarction, lethal or non-lethal stroke or lethal coronary heart disease), and of last recorded observation (follow-up or death), to define binary labels for 10-year ASCVD outcome and censoring. Individuals whose last recorded observation happened before an ASCVD event and before year 10 are considered censored. We remove records with extreme values of systolic blood pressure (outside 90–200 mm Hg), total cholesterol and HDL cholesterol (outside 130–320 and 20–100 mg/dL, respectively) or missing covariates.

Models
Unconstrained model
The original PCEs consist of four separate Cox proportional hazards models, stratified by sex and race, to account for differences in ASCVD incidence across the four groups (Black women, white women, Black men and white men).8 One revision of the PCEs, which reduced overfitting and improved calibration, replaced the Cox models with censoring-adjusted logistic regression models, stratified by sex and included race as a variable in each model.11 Our implementation of the unconstrained (UC) models consists of a single inverse probability of censoring (IPCW)-adjusted logistic regression model,23 and includes race and sex as binary variables. Censoring weights are obtained from four group-level Kaplan-Meier estimators applied to the training set. We include all features and their two-way interactions.

Group-recalibrated model
For recalibration, we logit-transform the predicted probabilities generated by the UC model, and use IPCW-adjusted logistic regression to fit a calibration curve for each group. We then use the resulting group-recalibrated model to obtain a set of recalibrated predictions.

EO model
The EO criterion requires that both the FPR and FNR be equal across groups at one or more thresholds.3 We use an in-processing method for constructing EO models,22 24 which provides a better calibration-EO tradeoff than the post-processing approach.25 We define the training objective by adding a regulariser to the UC model’s objective (online supplemental file A), with the degree of regularisation controlled by λ. The regulariser penalises differences between FPR and FNR at specified decision thresholds (7.5% and 20%), across the four groups.

Training procedures
Using random sampling stratified by group, outcome and presence of censoring, we divide our cohort into the training (80%), recalibration (10%) and test (10%) sets. Using the same procedure, we divide the training set into 10 equally-sized subsets and, for each subset, train a logistic regression model using stochastic gradient descent for up to 200 iterations of 128 minibatches, with learning rate of 10⁻⁴ on the remaining subsets. We terminate training if the cross-entropy loss does not improve on the held-out subset for 30 iterations. This procedure generates 10 UC models. To generate group-recalibrated models, we first generate predictions on the recalibration set, using the UC models (figure 2) and then use those train logistic regression models using BFGS (Broyden-Fletcher-Goldfarb-Shanno) optimisation, implemented in Scikit-Learn,26 with up to 10⁵ iterations. To examine the impact of the EO penalty, we repeat the unconstrained training procedure using the regularised training objective with four different settings of the parameter λ, distributed log-uniformly on the interval 0.1–1.0 (0.100,
0.215, 0.464, 1.000) and refer to the resulting models as EO1 through EO4. PyTorch V1.5.0 is used to define all models and training procedures. We make our code available at https://github.com/agataf/fairness_eval_ascvd.

**Evaluation**

We introduce threshold calibration error (TCE), a measure of local calibration, defined as the difference between the therapeutic threshold \( t_1=7.5 \% \) or \( t_2=20\% \) applied on the risk estimate and the implied threshold on the risk, measured by the calibration curve (figure 1C). As in the recalibration procedure, we estimate implied thresholds \( g(t) \) at a fixed therapeutic threshold \( t \) by fitting a calibration curve \( g \) for each group \( a \) (figure 1C). Then, for each threshold \( i \) we obtain TCE(i,a):

\[
TCE(i,a) = t_i - g_a
\]

A negative TCE indicates risk underestimation, since the threshold applied to the risk score is lower than the observed incidence of the outcome at that predicted risk level. Similarly, a positive TCE indicates risk overestimation.

To understand the tradeoff between TCE, FPR and FNR, we calculate intergroup SD (IGSD) between the four group-specific values of the three metrics. For a threshold \( i \), metric \( M \) and \( A \) distinct groups, IGSD\(_{M,i}\) is defined as:

\[
IGSD(M, t_i) = \sqrt{\frac{\sum_{a=1}^{A} (M_a - \mu_{M_i})^2}{A}}, \text{ where } \mu_{M_i} = \frac{\sum_{a=1}^{A} M_a}{A}
\]

IGSD captures the degree of performance disparity between groups; high IGSD in FPR and FNR corresponds to an EO disparity, and high IGSD in TCE corresponds to a treatment rule disparity.

For each of the four subgroups, and overall population, we report calibration and discrimination metrics at both the aggregate (absolute calibration error, ACE\(_{22}\) and area under the receiver operating characteristic, AUROC) and the threshold level (TCE, FPR and FNR) at \( t_1 \) and \( t_2 \), for the UC model, the group-recalibrated model (rUC) and the best-performing EO model, as well as the original PCE\(_{9}\) (PCE) and revised PCE\(_{11}\) (rPCE). We draw 1000 bootstrap samples from the test set, stratified by group and outcomes, to derive point estimates and 95% CIs for each metric. The 95% CIs are defined as the 2.5% to 97.5% percentiles of the distribution obtained via pooling over both the bootstrap samples and the 10 model replicates derived from the training procedure. We also report IGSD between the four group-specific median values in TCE, FPR and FNR at both thresholds. All metrics are computed over the uncensored population and adjusted for censoring using IPCW.

**RESULTS**

We describe the study population and present performance of the models. We report the TCE, FPR and FNR in figure 3, and IGSD of those three metrics in figure 4. We present results for EO3 in figure 3, as it was the only equalised odds model that achieved a reduction of IGSD (FPR) while keeping a low IGSD (FNR) at both thresholds. Results for the remaining EO models are included in online supplemental file B.

**Study population**

Overall, 25 619 individuals met the inclusion criteria, of whom 80% (N=20495) were assigned to the training set, and 10% (N=2562) to each the recalibration and test set. Table 1 summarises the mean age, ASCVD event incidence and frequency of censoring across the six data sets and four demographic groups. A cohort construction flowchart is included in online supplemental figure B1.

**Model performance**

The UC model achieved an overall AUROC of 0.827, (95% CI=0.800 to 0.853); comparing favourably with PCE (0.808 (0.779 to 0.835)) and rPCE (0.804 (0.777 to 0.831)), while maintaining differences of AUROC between groups (figure 3A). While UC had a slightly higher overall ACE (0.011 (0.006 to 0.023)) than rPCE (0.005 (0.001 to 0.015)), as well as a slightly higher local miscalibration at \( t_1 \) (TCE(t1) 0.012 (0.006 to 0.019) versus 0.000 (−0.004 to 0.005)), IGSD(TCE, \( t_1 \)) and IGSD(TCE, \( t_2 \)) both reduced under UC (from 0.018 to 0.004, and 0.053 to 0.016, respectively) (figure 4).

The group recalibration procedure (rUC) reduced the magnitude of TCE(T1) overall (−0.001 (−0.007 to 0.006)), and for each group, relative to UC (0.012 (0.006 to 0.019), table 1).

**Open access**
to 0.019)) (figure 3A). While recalibration improved TCE(t₂) overall (0.00 (−0.019 to 0.016) vs 0.006 (−0.013 to 0.023)), it increased the magnitude of miscalibration of individual groups—for instance, shifting TCE(t₂) from 0.033 (−0.016 to 0.066) to −0.071 (−0.196 to 0.008) for Black men, and increasing IGSD(TCE, t₂) to 0.038.

We also observe that, while TCE(t₁) and IGSD(TCE, t₁) improved for rUC, IGSD(FNR, t₁) worsened, increasing from 0.020 to 0.038, as did IGSD(FPR, t₂) and IGSD(FNR, t₂) (figure 4). Additionally, at each threshold, for all models, we observe a relationship between TCE, FPR and FNR: increased TCE (overestimation) leads to higher FPR and lower FNR, and decreased TCE (underestimation)—to lower FPR an higher FNR (figure 3B).

The EO procedure generated models with FPR and FNR which approached similar values across groups at both t₁ or t₂—bringing IGSD(FPR, t₁) to 0.054 from 0.094 (figure 4) while maintaining almost identical AUROC to UC (0.822 (0.794, 0.8)) (figure 3A). However, it did so by trading off error rates in opposite directions at the two thresholds, as described above (figure 3B). It also increased the magnitude of TCE at both thresholds (from 0.012 (0.006 to 0.019) to 0.033 (0.025 to 0.047) at t₁ and from 0.006 (−0.013 to 0.023) to −0.236 (−0.335 to −0.129) at t₂), and increased IGSD(TCE, t₁) to 0.011 (from 0.004) and IGSD(TCE, t₂) to 0.039 (from 0.016), implying that the scores generated by the EO model did not closely correspond to their calibrated values.

**DISCUSSION**

We identified local calibration of 10-year ASCVD risk prediction models at guideline-recommended thresholds as necessary for fair shared decision-making about statin treatment between patients and physicians. We find that the rPCEs differ in local calibration between groups—making guideline-compatibility of rPCE inconsistent across groups. We note that global measurements of calibration, used previously to evaluate the PCEs, did not capture this difference, illustrating the importance of local calibration evaluation.

Recalibrating the model separately for each group increased compatibility with guidelines at low levels of risk, while increasing intergroup differences in error rates. Conversely, estimators learnt with an EO constraint would not be concordant with existing guidelines as a result of induced miscalibration. Thus, absent a contextual analysis, fairness approaches that focus on error rates can produce misleading results.

In our experiments, group-recalibration did not improve calibration at t=20%. This may be due to the small sample size of the recalibration set, as well as of individuals predicted to be at high risk. This suggests that group-recalibration may not always be desirable, especially if local calibration of the UC model is deemed acceptable. However, improvement in local calibration observed at t=7.5% may be more relevant than calibration at higher risk levels for informing statin initiation decisions, since benefits of treatment are clearer at higher-risk levels.

Several design choices may have impacted the results, including the use of a single model with race and sex as variables in the UC and EO models, the use of a logistic regression as a recalibration method, and the use of an in-processing method that focused on particular decision.
thresholds to impose EO. We anticipate that alternative modelling choices would impact the size of the observed effects, but would likely not change the conclusions, since known statistical tradeoffs exist between EO and calibration.\textsuperscript{18,21,22}

Given this analysis, we recommend that developers building models for use with the ACC/AHA guidelines prioritise calibration across a relevant range of thresholds, and report group-stratified evaluation of local calibration alongside metrics of global fit. Before a model is deployed in a new setting, we recommend that it be evaluated on the target population, stratified by relevant groups—and group-recalibrated, if necessary. Knowledge about local miscalibration should also be incorporated into risk calculators to actively inform the physician-patient shared decision-making conversations, but should not replace recalibration efforts, since calibrated predictions are better suited for reasoning about potential consequences of treatment.\textsuperscript{20} 

Our analysis inherits the assumptions about relative importance of relevant risks and benefits used to derive therapeutic thresholds (online supplemental file C), which often fail to consider the impact of social determinants of health on treatment efficacy and of structural forms of discrimination in generating health disparities.\textsuperscript{26} Additionally, our use of self-identified racial categories—which can be understood as proxies for systemic and structural racist factors impacting health—may be inappropriate, potentially exacerbating historical racial biases and disparities in the clinical settings.\textsuperscript{29,30} Derivation of new risk prediction models may be necessary for multiethnic populations.\textsuperscript{12} Future work should explore decision analysis and modelling choices that incorporate this context.

CONCLUSION

Our analysis is one of the first to consider algorithmic fairness in the context of clinical practice guidelines. It illustrates general principles that can be used to identify contextually relevant fairness evaluations of models used in clinical settings in the presence of clinical guidelines. Such analysis should include careful consideration of the interplay between model properties, model-guided treatment policy, as well as the potential harms and benefits of treatment, for each relevant subgroup. At the same time, we note that striving for model fairness is unlikely to be sufficient in addressing health inequities, especially when their sources lay upstream of the model-guided intervention, as is the case of structural racism.\textsuperscript{30} We encourage future work to situate fairness analyses in this broader context.

Acknowledgements We thank Scotty Fleming, Alison Callahan, Sanjana Srivastava, Sherri Rose, Jonathan Lu, Ethan Steinberg, Yizhe Xu, Jennifer Wilson, Nikolaos Ignatiadis, Steve Yadlowsky and Julian Genkins for insightful comments.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Cohort characteristics for patients who met inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Black women</td>
</tr>
<tr>
<td>ARIC</td>
<td>1812</td>
</tr>
<tr>
<td>CARDIA</td>
<td>232</td>
</tr>
<tr>
<td>CHS</td>
<td>304</td>
</tr>
<tr>
<td>JHS</td>
<td>1310</td>
</tr>
<tr>
<td>MESA</td>
<td>768</td>
</tr>
<tr>
<td>All</td>
<td>4426</td>
</tr>
<tr>
<td></td>
<td>Non-Black women</td>
</tr>
<tr>
<td>ARIC</td>
<td>4815</td>
</tr>
<tr>
<td>CARDIA</td>
<td>289</td>
</tr>
<tr>
<td>CHS</td>
<td>1848</td>
</tr>
<tr>
<td>FHS OS</td>
<td>828</td>
</tr>
<tr>
<td>MESA</td>
<td>1913</td>
</tr>
<tr>
<td>All</td>
<td>9693</td>
</tr>
</tbody>
</table>

Data are grouped by sex and race, as well as data set. Each group of patients is described by four values: total number of individuals, mean age, censoring-adjusted incidence of ASCVD events within 10 years of the initial examination and fraction of censored individuals.

ASCVD event incidence was calculated by weighing the number of positive outcome and negative outcome uncensored individuals with the sum of their inverse probability of censoring weights.

ARIC, Atherosclerosis Risk in Communities Study; ASCVD, atherosclerotic cardiovascular disease; AUROC, area under the receiver operating characteristic; CARDIA, Coronary Artery Risk Development in Young Adults Study; CHS, Cardiovascular Health Study; FHS OS, Framingham Heart Study Offspring Cohort; FNR, false negative rate; FPR, false positive rate; IPCW, inverse probability of censoring; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis; PCE, Pooled Cohort Equations.
and discussion. Approval for this non-human subjects research study is provided by the Stanford Institutional Review Board, protocol IB-46829.

Contributors AF and SRP conceived of the present idea, and developed it with support from BP and NS. AF cleaned and preprocessed the data, trained predictive models, generated and analysed model evaluations. SRP provided code for training and evaluation and analysed model evaluations. AF drafted the manuscript, and revised it with SRP in consultation with BP and NS. NS supervised the project. AF is the guarantor of the project.

Funding This study was funded by Stanford University School of Medicine (Program for AI in Healthcare); National Heart, Lung, and Blood Institute (R01 HL144555).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The six data sets used: ARIC (Atherosclerosis Risk in Communities Study), CARDIA (Coronary Artery Risk Development in Young Adults Study), CHS (Cardiovascular Health Study, 1989–1999), FHS OS (Framingham Heart Study Offspring Cohort), MESA (Multi-Ethnic Study of Atherosclerosis) and JHS (Jackson Heart Study), are available on BioLINCC.

Supplemental material This content has been supplied by the author(s).

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and to use the work for any purpose — for non-commercial purposes the original author(s) and source must be credited.

ORCID iDs Agata Foryciarz http://orcid.org/0000-0002-8968-5805 Nigam Shah http://orcid.org/0000-0001-9393-7158

REFERENCES

1 Regularization approaches for equalized odds at a threshold

Let \( D = \{(x_i, y_i, a_i)\}_{i=1}^N \sim P(X, Y, A) \) be a dataset, where \( X \) designates patient-level features, \( Y \) is a binary indicator for the occurrence of an outcome, and \( A \) is a discrete attribute that stratifies the population into \( K \) disjoint groups. We learn a model \( f_\theta(x) \) to estimate \( E[Y \mid X = x] = P(Y = 1 \mid X = x) \) using empirical risk minimization:

\[
\min_{\theta \in \Theta} \mathbb{E}_{(x,y) \sim D} \ell(y, f_\theta(x)),
\]

(1)

where \( \ell \) is the cross-entropy loss and the expectation with respect to the dataset \( D \) indicates an empirical average over the dataset.

To construct a training objective that penalizes violation of equalized odds, we add a regularizer \( M_{EqOdd} \) to the training objective, with the degree of regularization controlled by a non-negative scalar \( \lambda \):

\[
\min_{\theta \in \Theta} \mathbb{E}_{(x,y) \sim D} \ell(y, f_\theta(x)) + \lambda M_{EqOdd}.
\]

(2)

We derive a differentiable regularizer for this setting by first noting that the true positive rate at a threshold \( t \) can be defined as

\[
TPR := \mathbb{E}_{x \sim D \mid Y = 1} h(f_\theta(x) - t),
\]

(3)

where \( \mathbb{E}_{(x,y) \sim D \mid Y = 1} \) indicates the expectation over the subset of the population for which \( Y = 1 \), and \( h(z) = 1\{z > 0\} \) is the step function. Similarly, the false positive rate is given by

\[
FPR := \mathbb{E}_{x \sim D \mid Y = 0} h(f_\theta(x) - t).
\]

(4)

In order to achieve equal sensitivity (true positive rate) across groups and equal specificity (true negative rate) across groups, it is sufficient for the model to achieve a true positive rate and false positive rate for each group equal to the true positive rate and false positive rate of the population overall. The direct incorporation of the true positive rate and false positive rate, as formulated above, does not provide a useful signal for stochastic gradient descent when incorporated into a regularizer because the step function \( h(z) \) has a derivative of zero everywhere that the derivative is defined. To address this issue, we replace the step function \( h(z) \) with the sigmoid function \( \sigma(z) = \frac{1}{1+\exp(-z)} \), following Cotter et al. [1]. Furthermore, in practice, we compare the log of the scores to the log of the threshold. Terms representing the difference between the relaxed rates for a group \( A_k \) with the population are given by:

\[
M_{TPR}^{k,t} = \mathbb{E}_{x \sim D \mid A = A_k, Y = 1} \left[ \sigma(\log f_\theta(x) - \log(t)) \right] - \mathbb{E}_{x \sim D \mid Y = 1} \left[ \sigma(\log f_\theta(x) - \log(t)) \right]
\]

(5)
and
\[ M_{FPR}^{k,t} = \mathbb{E}_{x \sim D | A = A_k, Y = 0} \left[ \sigma(\log f_\theta(x) - \log(t)) \right] - \mathbb{E}_{x \sim D | Y = 0} \left[ \sigma(\log f_\theta(x) - \log(t)) \right]. \]  

(6)

To construct the regularizer, we combine these terms into a single non-negative regularizer over all groups and thresholds:

\[ M_{EqOdd} = \sum_{t_i \in \{t_1, t_2\}} \sum_{k=1}^{K} \left[ (M_{TPR}^{k,t_i})^2 + (M_{FPR}^{k,t_i})^2 \right]. \]  

(7)

To adjust for censoring, we use an inverse probability of censoring (IPCW) approach, and replace the empirical averages in equations 1, 2, 5, and 6 with weighted ones that incorporate the IPCW weights.

**References**

Supplementary Figure B1. Cohort construction flowchart. We excluded individuals younger than 40 or older than 79, those with past history of myocardial infarction, stroke, coronary bypass surgery or angioplasty, congestive heart failure or atrial fibrillation, and those who were receiving statins at the time of the initial exam. We removed records with extreme values of systolic blood pressure (outside 90-200 mmHg), total cholesterol and high-density lipoprotein cholesterol (outside 130-320 and 20-100 mg/dL, respectively) and those with missing features.
Supplementary Figure B2. Performance of equalized odds models with different settings $\lambda$, stratified by demographic group, and evaluated on the held out test set. EO1 - Equalized Odds ($\lambda=0.1$), EO2 - Equalized Odds ($\lambda=0.215$), EO - Equalized Odds ($\lambda=0.464$), EO4 - Equalized Odds ($\lambda=1$). The left panel shows AUROC and Absolute Calibration Error. The middle panel shows Threshold Calibration Error at two therapeutic thresholds (7.5% and 20%). The right panel shows False Negative and False Positive Rates at the two therapeutic thresholds.
THE MEANING OF THRESHOLDS

The American College of Cardiology and the American Heart Association (ACC/AHA) 2019 guidelines on primary prevention of atherosclerotic cardiovascular disease (ASCVD) recommend different statin treatment regimens based on an estimate of ASCVD risk paired with therapeutic thresholds of 7.5% and 20% [1,2]. When grounding fairness analysis in threshold-based clinical guidelines, it is important to consider the assumptions underlying those thresholds, which can have implications for fairness.

The 7.5% treatment threshold for individuals without diabetes or clinical ASCVD and non-elevated LDL-C was established as a risk level above which net absolute benefit is positive. This was based on evidence from multiple RCTs, which studied the impact of statin treatment on ASCVD risk reduction [2] (Supplementary Section 7.3). The guidelines quantified net absolute benefit as a difference between benefits and harms, with benefit defined as the number of patients who would need statin treatment to prevent one ASCVD event over 10 years (number needed to treat, NNT), and harm the number of statin-treated patients who would yield one excess case of diabetes over 10 years (number needed to treat to harm, NNH), both of which depend on 1) the estimate of a 10-year ASCVD risk of a patient, as well as 2) risk reduction from statin dose.

In 2013, The ACC/AHA committee [2] considered 5% and 7.5% as therapeutic thresholds. Those numbers corresponded to 10-year prevalence of ASCVD in the control groups of relevant primary prevention RCTs (Air Force/Texas Coronary Atherosclerosis Prevention Study [AFCAPS/TEXCAPS], Collaborative Atorvastatin Diabetes Study [CARDS], Justification for the Use of Statins in Primary Prevention: An Intervention Trial Evaluating Rosuvastatin [JUPITER], and Management of Elevated Cholesterol in the Primary Prevention Group of Adult Japanese [MEGA]). The NNT and NNH for different statin doses were then compared at the thresholds. At 7.5%, NNH was higher than NNT for both moderate and high-intensity statin treatment - suggesting that treating patients with either dose would prevent more ASCVD events in more people than it would induce new cases of diabetes in. This observation led the committee to select 7.5% as an acceptable therapeutic threshold.

This threshold is not derived from an economic sense of utility, where the relative value of harms and benefits are accounted for. While the guidelines explicitly state that the harm of an ASCVD event is more significant than that of a new diabetes diagnosis, the relative weight of NNH and NNT is not quantified. If NNH and NNT were treated as equivalent, the optimal decision threshold would correspond to a risk value where NNH=NNT - and it would have been lower than 7.5% (~2.5% for moderate intensity statins, and ~6% for high intensity statins). The committee made their final recommendation based on the interpretation of the best available data and methods at the time. Therefore, the 7.5% threshold could be considered reflective of expert consensus rather than an algorithmic optimization.
The 20% threshold was added in the 2018 Cholesterol Clinical Practice Guidelines (Section 4.4.2) [1] as a risk value above which high-intensity statin therapy should be considered (with 7.5-20% range considered more appropriate for considering moderate- or high-intensity therapy). Further analysis of trial outcomes conveyed a relationship between proportional reduction in LDL cholesterol from baseline and reduction in ASCVD. Thus, a higher threshold was meant to convey the importance of higher intensity therapy to provide maximal benefit for high risk patients. This was enacted by introducing an additional decision threshold in the guidelines. Additionally, the 2019 Guidelines [5] included LDL-C reduction goals, rather than statin doses, for the different risk categories.

While we have heard anecdotal evidence in our conversations with physicians which implied that the thresholds were set to account for miscalibration, the guidelines imply that thresholds were not set explicitly to account for known miscalibration of the PCE models.

The guidelines also make several other assumptions, listed below, which are detailed in Table 5 of the Supplement to 2013 ACC/AHA guideline on the treatment of blood cholesterol [2].

- Similar relative risk reduction (RRR) for CVD events across patient groups.
- RRR is proportional to the degree of LDL-C lowering by statin therapy.
- Absolute benefit in risk reduction is proportional to baseline risk of group/individual and the intensity of statin therapy - patients and groups with higher absolute risk will get more absolute benefit from statins.
- Absolute risk for adverse outcomes is proportional to baseline risk of a given group, and the intensity of statin therapy.
- Similar relative risk reduction (RRR) for CVD events across patient groups.
- RRR is proportional to the degree of LDL-C lowering by statin therapy.
- Absolute benefit in risk reduction is proportional to baseline risk of group/individual and the intensity of statin therapy - patients and groups with higher absolute risk will get more absolute benefit from statins.
- Absolute risk for adverse outcomes is proportional to baseline risk of a given group, and the intensity of statin therapy.

Since their publishing, some of those assumptions have been reevaluated [6].
REFERENCES


