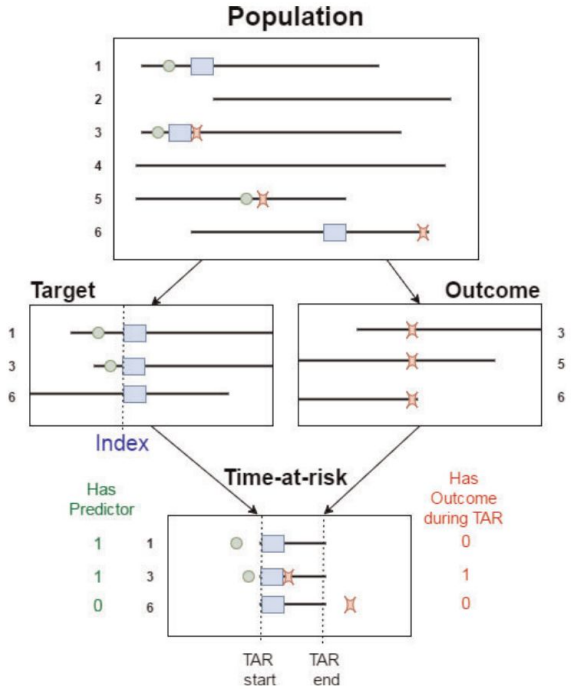
Recommendations for algorithmic fairness assessments of predictive models in healthcare: evidence from large-scale empirical analyses

Stephen Pfohl
PhD Candidate
Stanford Biomedical Informatics
November 12, 2021

Machine learning with electronic health records



Reps, Jenna M., et al. "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data." Journal of the American Medical Informatics Association 25.8 (2018): 969-975.

Aspirations for algorithmic fairness in healthcare

1. Assess systematic differences in model behavior or performance across patient populations for model reporting and auditing
2. Mitigate systematic differences in model behavior or performance across patient populations
3. Build models that predict outcomes well for each population
4. Proactively identify and mitigate *upstream biases* in data collection, problem formulation, and measurement
5. Ensure that machine-learning-enabled interventions prevent the exacerbation of disparities and promote health equity

Limiting the scope

1. Assess systematic differences in model behavior or performance across patient populations for model reporting and auditing
2. Mitigate systematic differences in model behavior or performance across patient populations
3. Build models that predict outcomes well for each population
4. Proactively identify and mitigate *upstream biases* in data collection, problem formulation, and measurement
5. Ensure that machine-learning-enabled interventions prevent the exacerbation of disparities and promote health equity

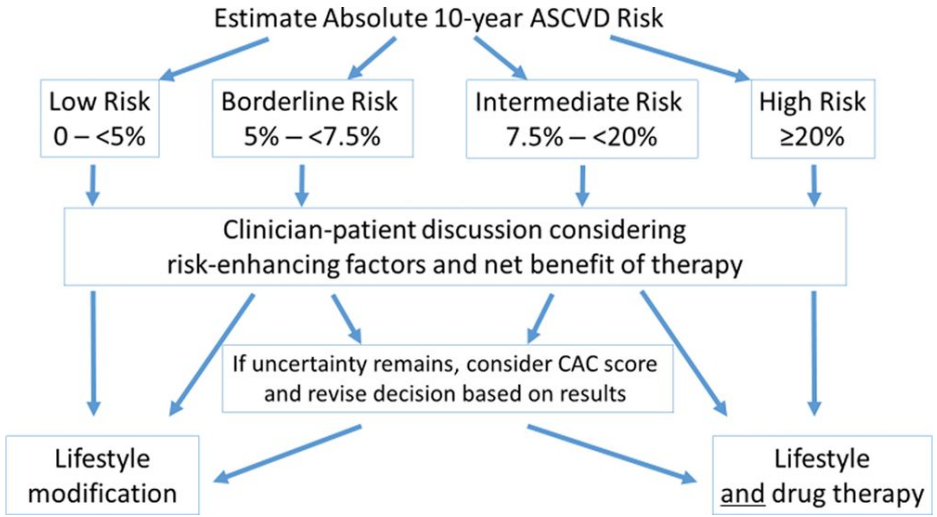
Questions explored in this work

- What are the best practices for conducting algorithmic fairness assessments to evaluate predictive models of clinical outcomes?
  - What should be measured and why?
  - How should the results be interpreted?
- What are the best practices for developing predictive models that enable fair clinical decision making?
  - Should model training objectives include explicit fairness constraints?
  - How can we build models that predict outcomes well for each group?

Relationship to papers and projects

1. **Pfohl, S.R,** Xu, Y., Foryciarz, A., Ignatiadis N., Genkins J., & Shah, N. H (2021). Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare.
2. Foryciarz, A., **Pfohl, S. R.**, Patel, B., & Shah, N.H. (2021). Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health and Care Informatics (In press).*
3. **Pfohl, S. R.**, Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., & Shah, N. H. (2021). A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *arXiv preprint arXiv:2108.12250*.
4. **Pfohl, S. R.**, Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113, 103621.
5. **Pfohl, S. R.**, Duan, T., Ding, D. Y., & Shah, N. H. (2019). Counterfactual reasoning for fair clinical risk prediction. *In Machine Learning for Healthcare Conference* (pp. 325-358). PMLR.
6. **Pfohl, S.R**., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., & Shah, N. H. (2019). Creating fair models of atherosclerotic cardiovascular disease risk. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 271-278).

Example: primary prevention of atherosclerotic cardiovascular disease (ASCVD)



Lloyd-Jones, *et al* . (2019). Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology. *Circulation*, *139*(25), E1162–E1177.

Constructing an ASCVD cohort

- Extract from claims database
  - All patients age 40-75, no prior CVD or statin prescription
- Define 10-year ASCVD as any
  - MI, stroke, or fatal CHD
- Define censoring event as earliest of
  - End of enrollment, statin prescription, or death
- Feature extraction
  - All prior conditions, procedures, lab orders (+abnormal flags), medications, and demographics

Group	Count	Censoring rate	Incidence
Female	3,253,609	0.816	0.105
Male	2,549,256	0.821	0.120
Asian	165,198	0.814	0.0829
Black	438,144	0.786	0.136
Hispanic	433,238	0.800	0.104
Other	880,116	0.936	0.115
White	3,886,169	0.797	0.110
Asian, female	88,100	0.806	0.0793
Asian, male	77,098	0.823	0.0874
Black, female	262,559	0.784	0.128
Black, male	175,585	0.788	0.150
Hispanic, female	235,736	0.792	0.102
Hispanic, male	197,502	0.810	0.107
Other, female	522,369	0.938	0.108
Other, male	357,747	0.932	0.125
White, female	2,144,845	0.794	0.102
White, male	1,741,324	0.802	0.119
Type 2 diabetes absent	5,388,193	0.817	0.104
Type 2 diabetes present	414,672	0.835	0.20
Type 1 diabetes absent	5,741,282	0.818	0.110
Type 1 diabetes present	61,583	0.825	0.240
RA absent	5,733,505	0.819	0.110
RA present	69,360	0.782	0.185
CKD absent	5,758,773	0.819	0.110
CKD present	44,092	0.767	0.253

Pfohl, S.R, *et. al*. (2021). Algorithmic fairness assessments for atherosclerotic cardiovascular disease risk estimation: calibration, net benefit, and equalized odds.

What are the best practices for conducting algorithmic fairness assessments to evaluate predictive models of clinical outcomes?

- What should be measured and why?
- How should the results be interpreted?

A prerequisite: transparent reporting and design

1. Identify intended intervention its effectiveness
2. Identify stakeholders, their values, and any conflicts
    a. Incl. patients and especially underrepresented or marginalized populations
3. Document dataset preparation, cohort construction, and model development protocols
4. Clearly specify and justify assumptions on data generating mechanisms and measurement processes
    a. Ideally, there is no unmodeled differential measurement error in outcomes across groups

Recommendations for evaluation

1. Report and contextualize stratified performance metrics
2. Prioritize calibration-based fairness assessments
3. Do not consider differences in TPR, FPR, PPV, or classification rates as being necessarily problematic
4. Do not consider context-free fairness assessments as sole indicators of whether ML-intervention introduces/exacerbates harm or is equity-promoting

Definitions and notation

- Dataset: $\mathcal{D} = \{x_i, y_i, a_i\}_{i=1}^{N} \sim P(X, Y, A)$
- Features: $X \in \mathcal{X} = \mathbb{R}^m$
- Binary outcome: $Y \in \mathcal{Y} = \{0, 1\}$
- Attribute stratifying the data into *K* groups: $A \in \mathcal{A} = \{A_k\}_{k=1}^{K}$
- Model: $f_\theta : \mathcal{X} \to [0, 1]$
- Score: $S = f_\theta(X)$
- Threshold-predictor: $\hat{Y} = \mathbb{I}[S \geq \tau]$
- Bayes-calibrated score: $f^*(x) = \mathbb{E}[Y \mid X = x]$
- Calibration curve: $c(s) = \mathbb{E}[Y \mid S = s]$

Summary of algorithmic fairness criteria

Criteria	Definition	Interpretation
Metric parity	$g(\cdot) \perp A$	Metric g does not differ
Demographic parity	$S \perp A$	Score distribution does not differ
Demographic parity	$\hat{Y} \perp A$	Classification rate does not differ
Equalized odds	$S \perp A \mid Y$	Identical ROC curves
Equalized odds	$\hat{Y} \perp A \mid Y$	Equal TPR and equal FPR
Group calibration	$\mathbb{E}[Y \mid S = s, A] = s$	Calibrated for each group
Sufficiency	$Y \perp A \mid S$	Calibration curves do not differ
Predictive parity	$Y \perp A \mid \hat{Y} = 1$	PPV does not differ

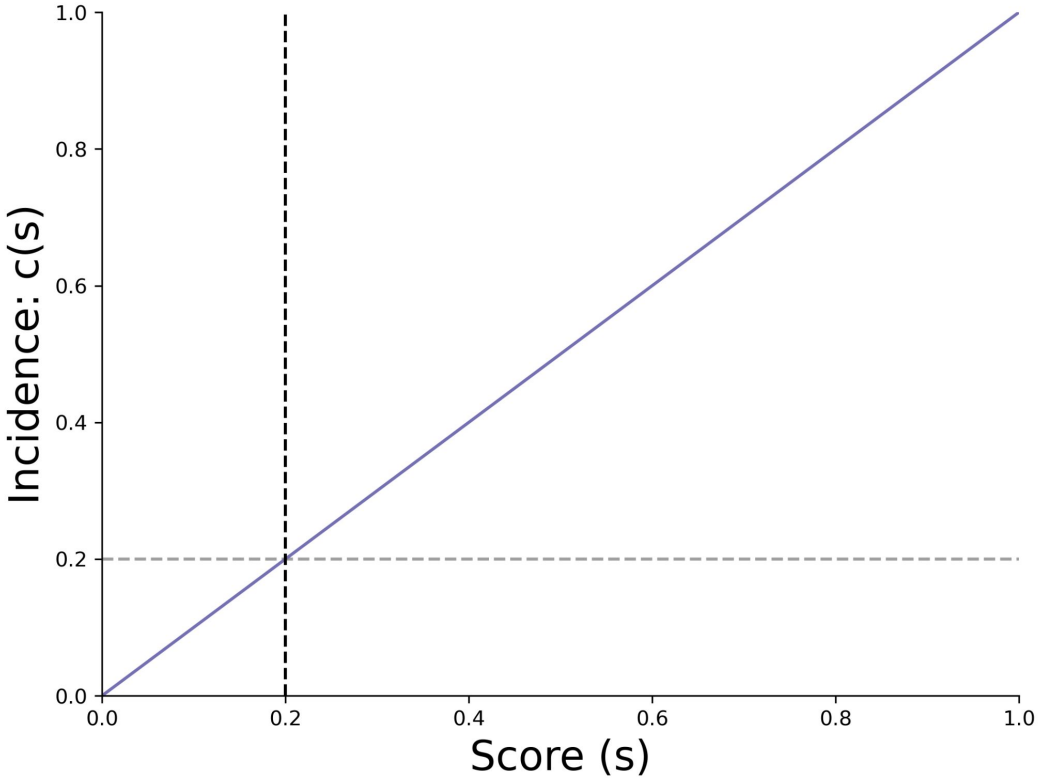
Calibration and sufficiency

Criteria	Definition	Interpretation
Metric parity	$g(\cdot) \perp A$	Metric g does not differ
Demographic parity	$S \perp A$	Score distribution does not differ
Demographic parity	$\hat{Y} \perp A$	Classification rate does not differ
Equalized odds	$S \perp A \mid Y$	Identical ROC curves
Equalized odds	$\hat{Y} \perp A \mid Y$	Equal TPR and equal FPR
Group calibration	$\mathbb{E}[Y \mid S = s, A] = s$	Calibrated for each group
Sufficiency	$Y \perp A \mid S$	Calibration curves do not differ
Predictive parity	$Y \perp A \mid \hat{Y} = 1$	PPV does not differ

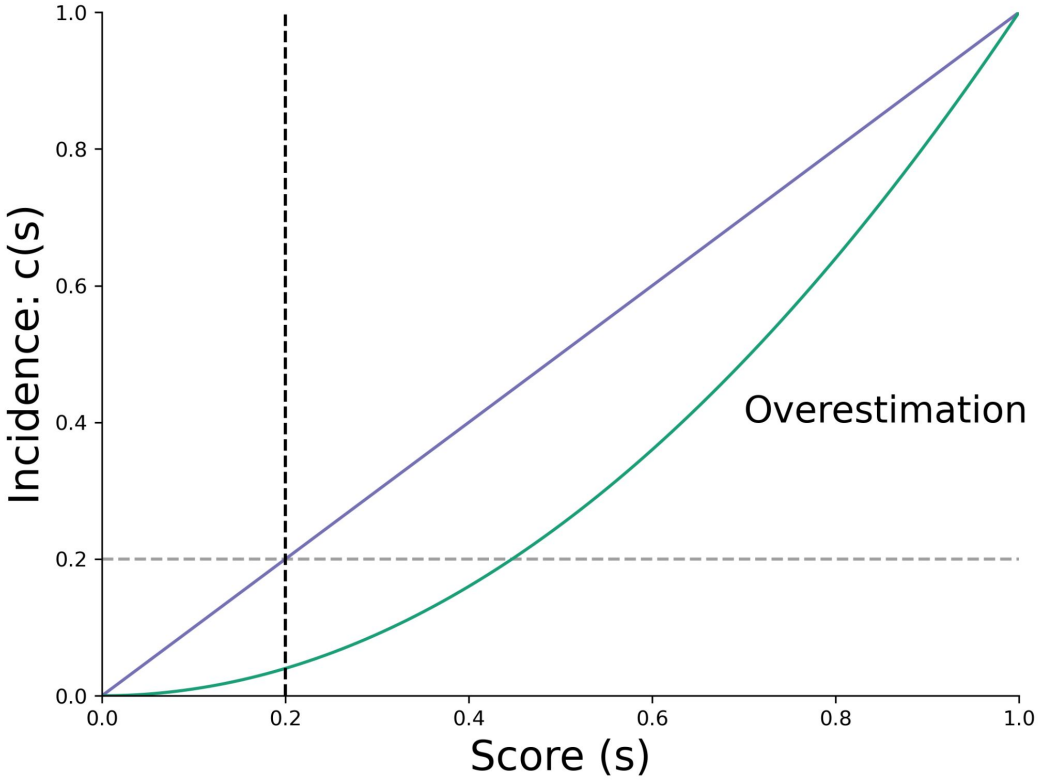
Why calibration?

- Calibration enables informed interpretation of risk estimates as probabilities
- Group calibration and sufficiency are consistent with the goal of treating the set of patients with a given risk of the outcome similarly across groups

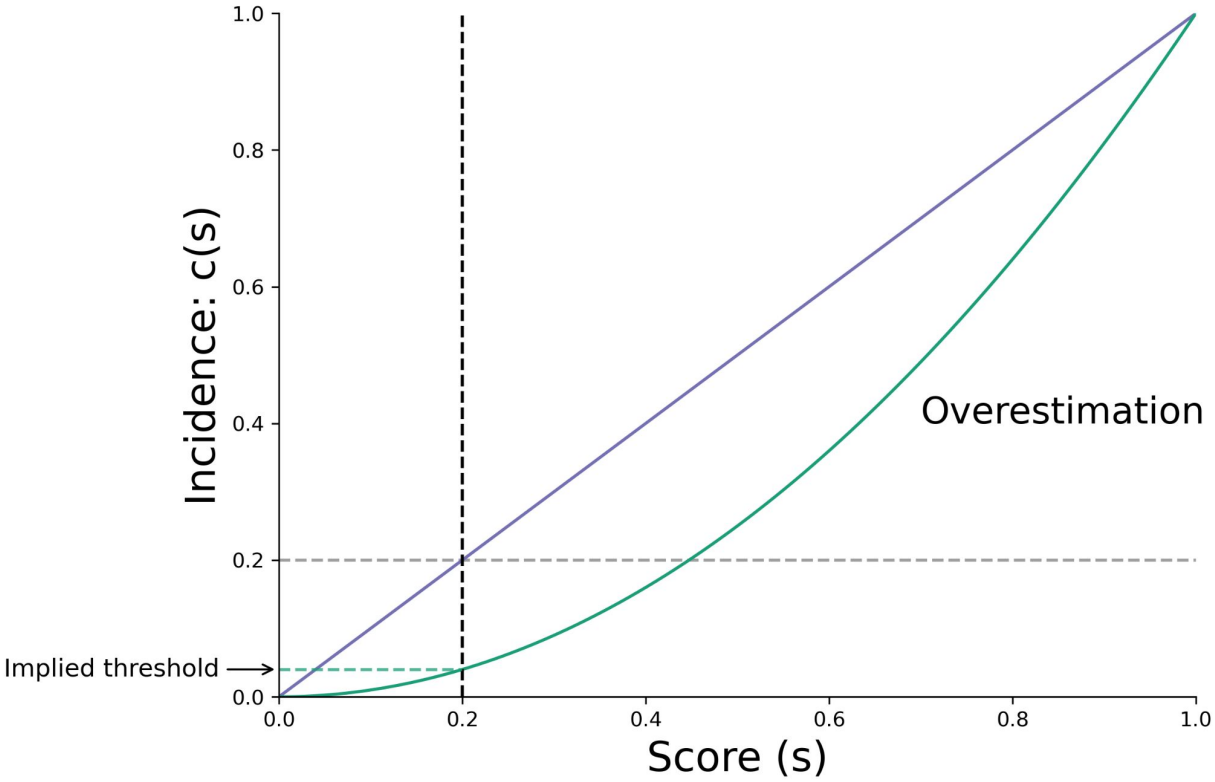
Calibration curves: perfect calibration



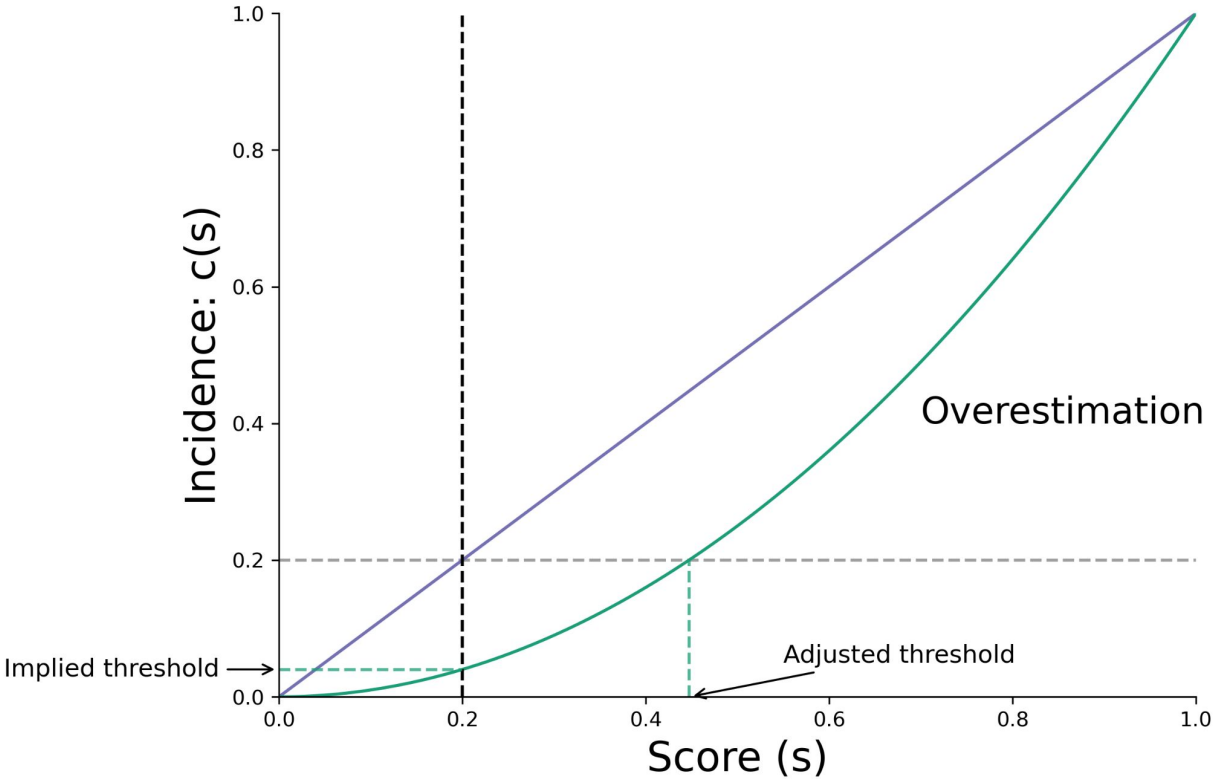
Calibration curves: assessing overestimation



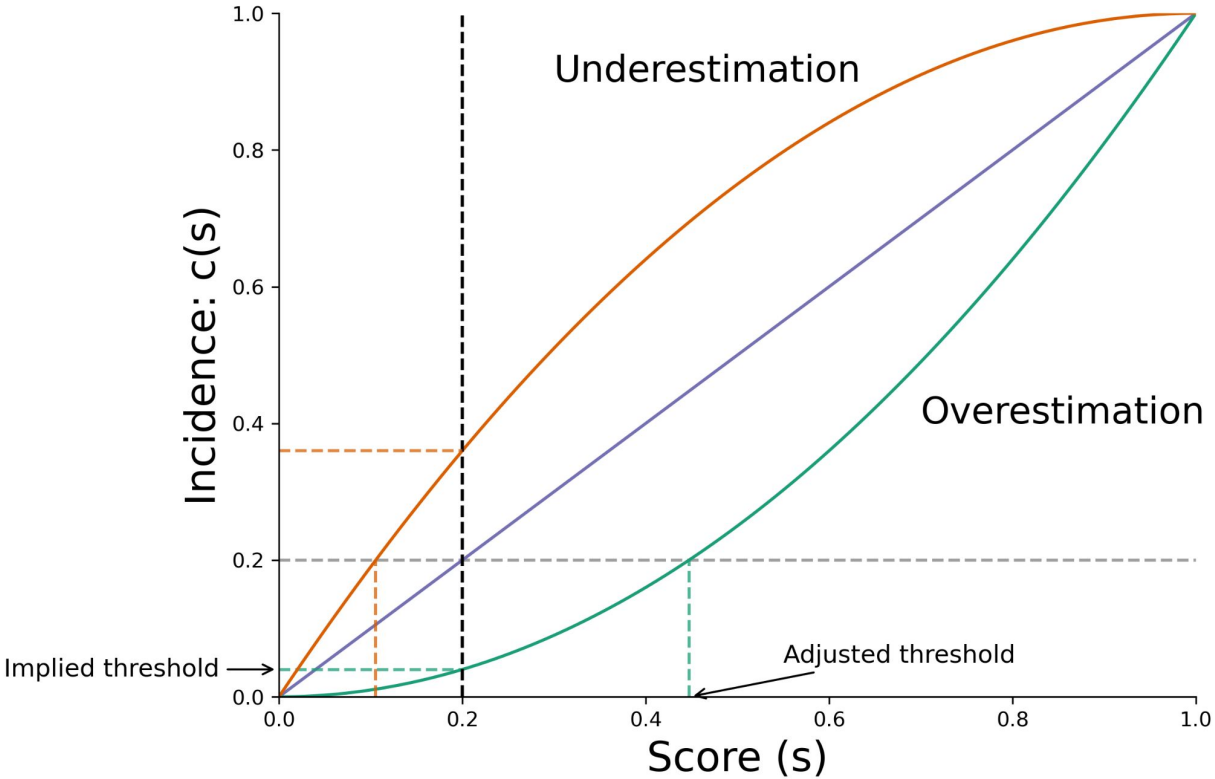
Calibration curves: implied threshold



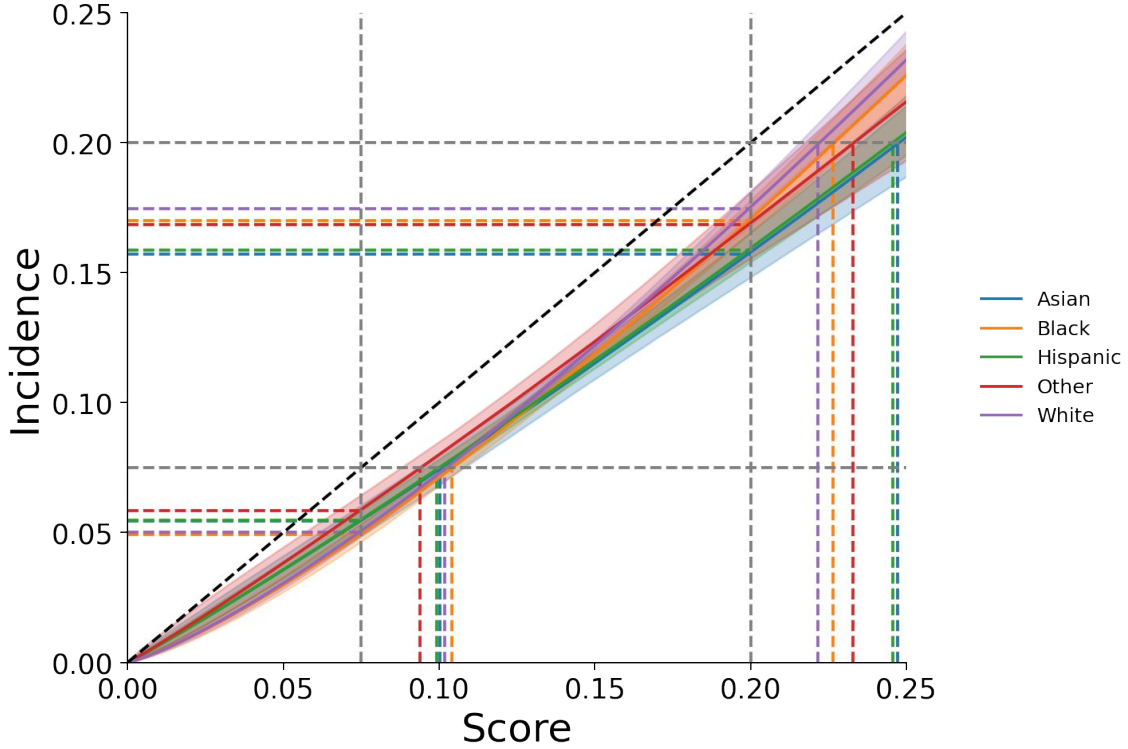
Calibration curves: adjusted threshold



Calibration curves: assessing over/underestimation



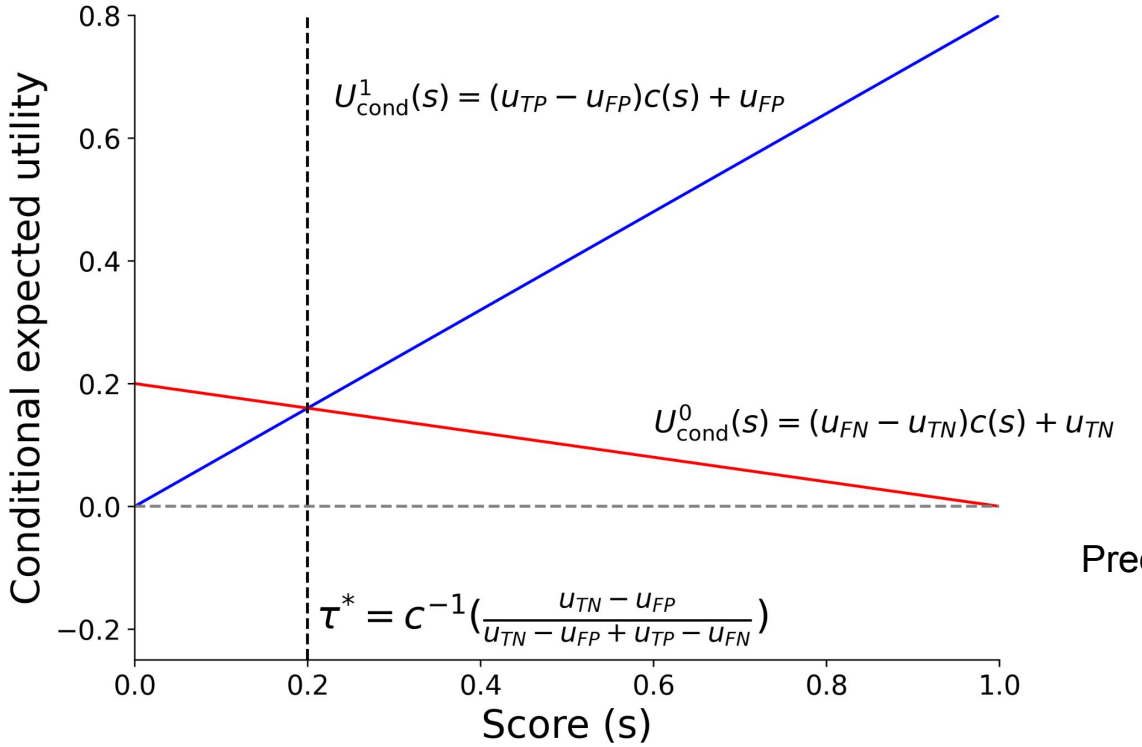
Calibration of a ten-year ASCVD risk estimator



Calibration, thresholds, and utility

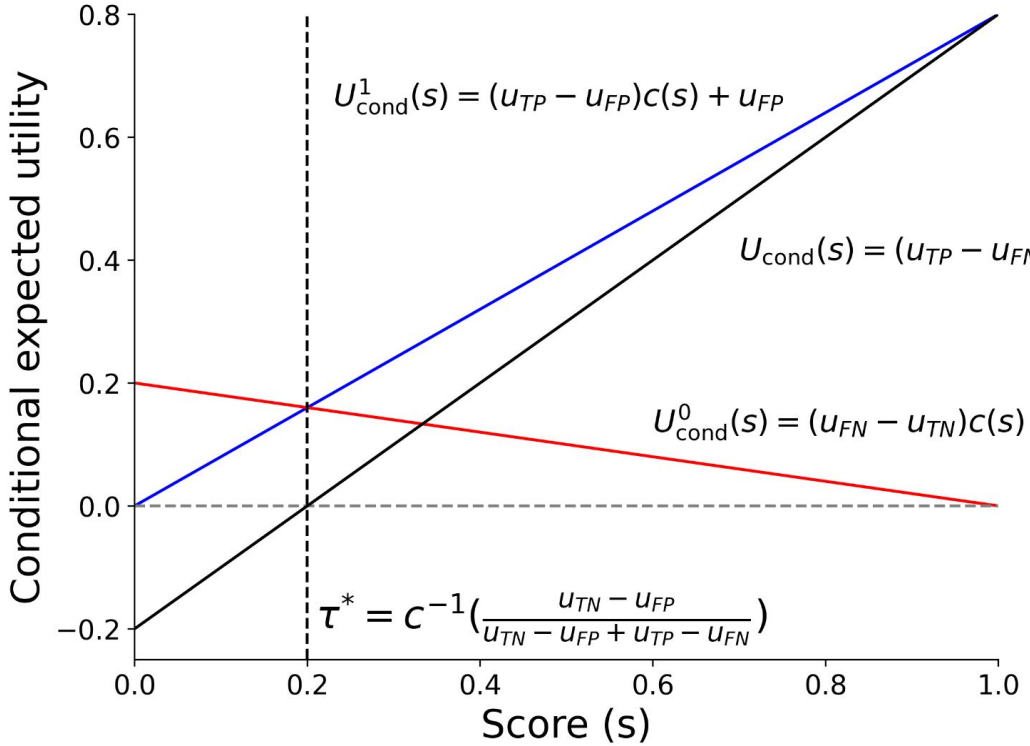
- The calibration curve can be used as a proxy for the *conditional utility* of the intervention conditioned on the risk score
- Sufficiency implies a shared utility-maximizing threshold if properties of the intervention and preferences do not differ across groups

Conditional utility and the fixed-cost utility function



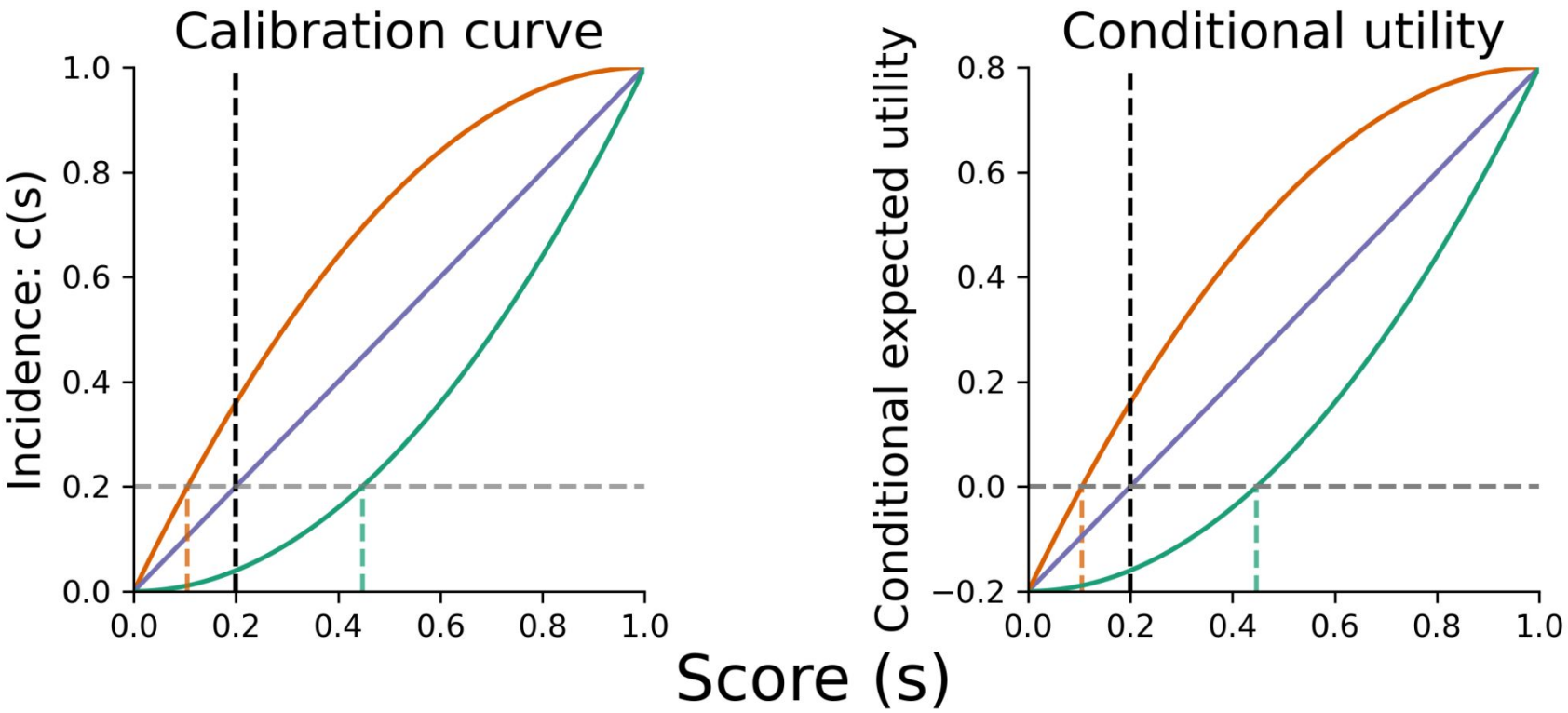
Prediction \ Outcome	+	-
+	u_{TP}=0.8	u_{FP}=0
-	u_{FN}=0	u_{TN}=0.2

Conditional utility and the fixed-cost utility function

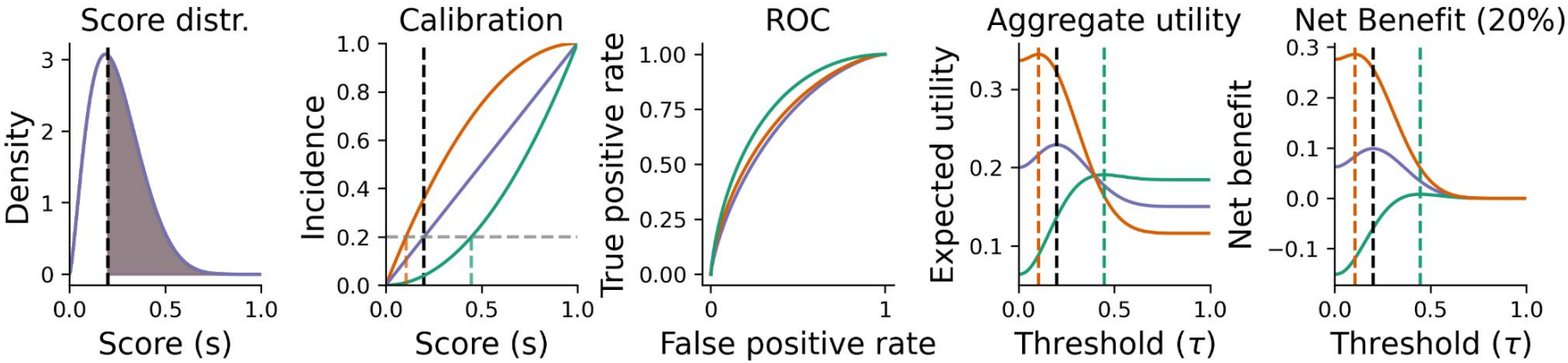


		Outcome +	Outcome -
Prediction	+	u_{TP}=0.8	u_{FP}=0
	-	u_{FN}=0	u_{TN}=0.2

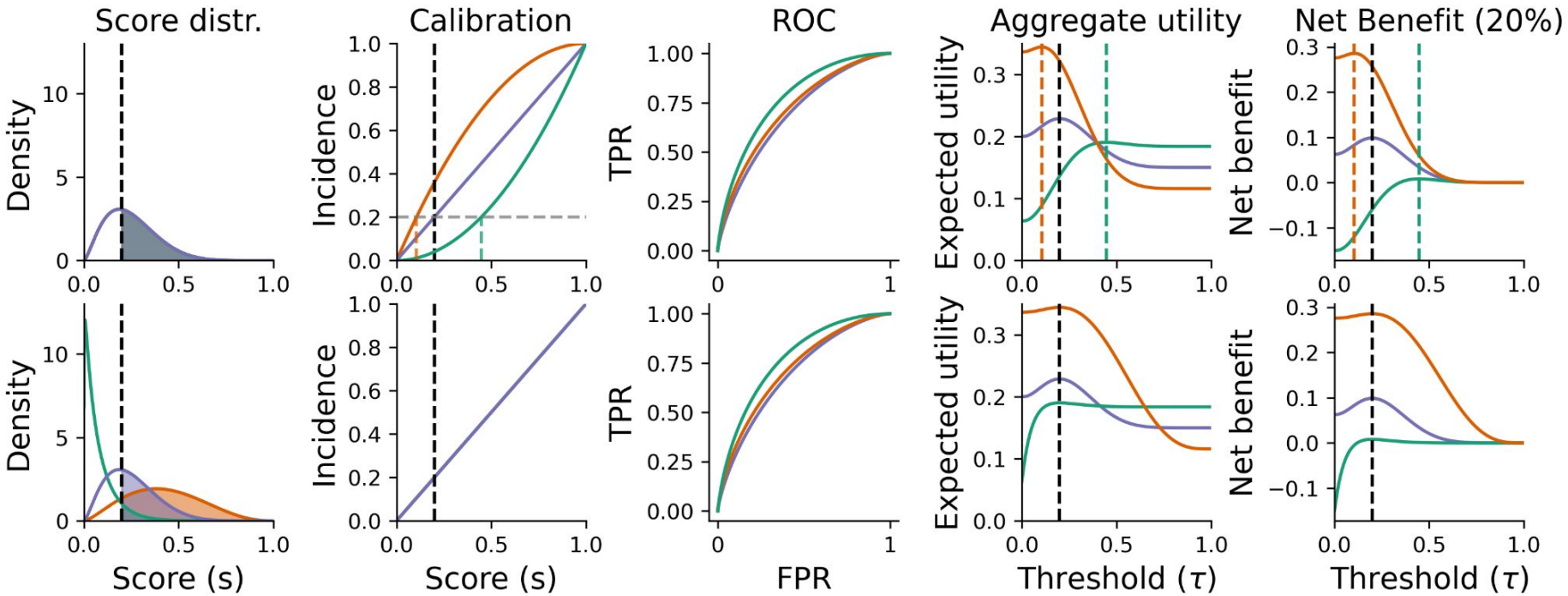
Calibration and conditional utility (simulation)



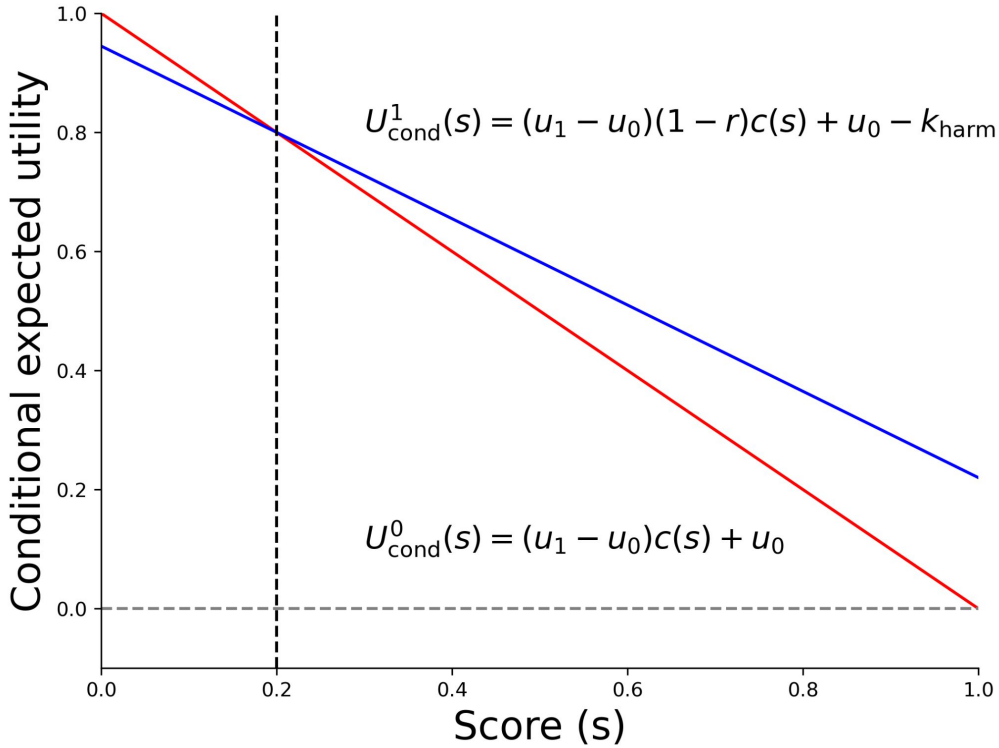
Effect of calibration on optimal threshold (simulation)



Calibrated models share a maximum utility threshold



Extending the approach for ASCVD risk estimation:
conditional utility with constant relative risk reduction



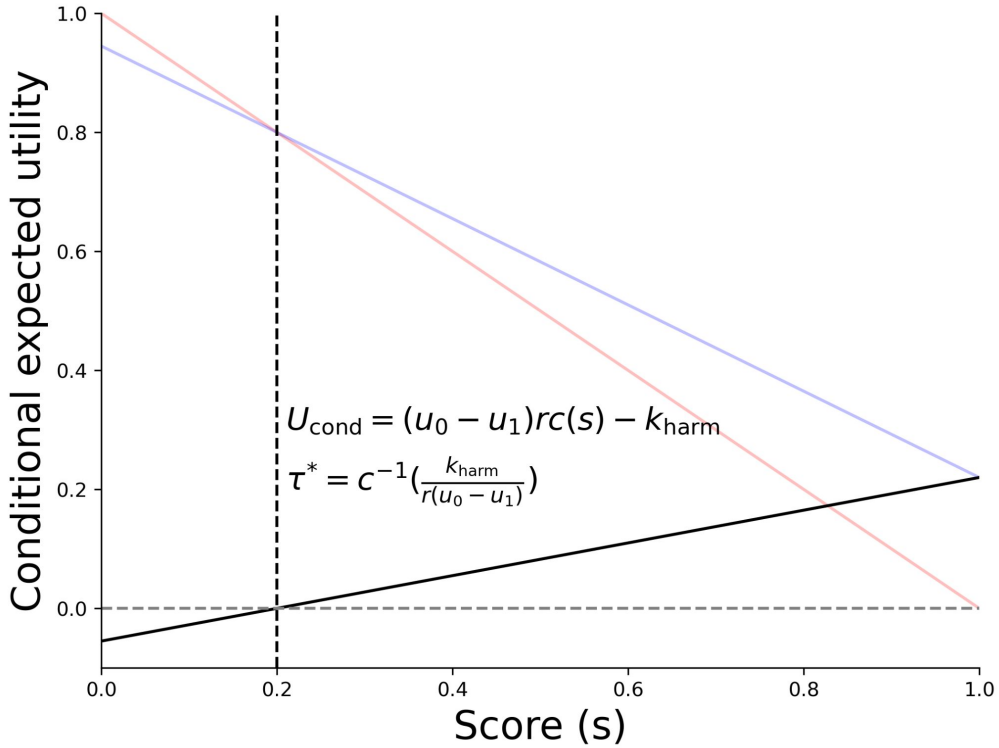
Utility of no ASCVD: u_0=1

Utility of ASCVD: u_1=0

Relative risk reduction: r=0.275

Expected harm: k_{harm}=0.055

Conditional utility with constant relative risk reduction



Utility of no ASCVD: u_0=1

Utility of ASCVD: u_1=0

Relative risk reduction: r=0.275

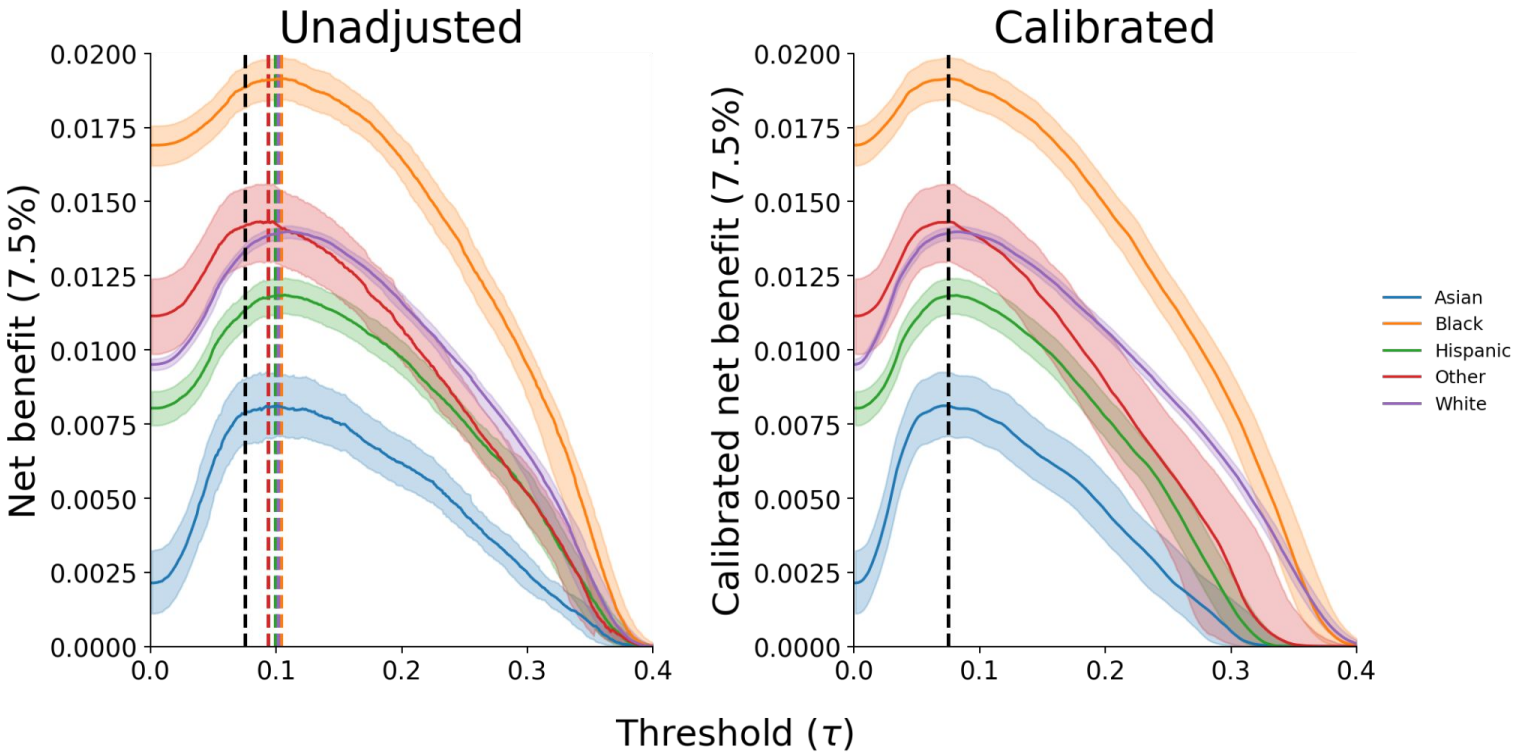
Expected harm: k_{harm}=0.055

Net benefit of statin initiation

- Net benefit assuming expected harm is independent of risk estimate
  - $\mathrm{NB}(\tau_y;\tau_y^*) = -\mathbb{E}[p_y^0(s) \mid S < \tau_y]P(S < \tau_y) - \mathbb{E}[p_y^1(s) \mid S \geq \tau_y]P(S \geq \tau_y) - \mathrm{ARR}(\tau_y^*)P(S \geq \tau_y) + P(Y = 1)$
- If relative risk reduction is constant
  - $\mathrm{NB}(\tau_y;\tau_y^*) = -(1 - \mathrm{NPV}(\tau_y))P(S < \tau_y) - P(S \geq \tau_y)\Big((1-r)\mathrm{PPV}(\tau_y) + r\tau_y^*\Big) + P(Y = 1)$
- Simplifying assumptions, following Soran *et. al* [1]
  - Use a constant expected relative risk reduction of 27.5%, assuming the use of moderate intensity statin (20 mg atorvastatin)

[1] Soran, H., Schofield, J. D., & Durrington, P. N. (2015). Cholesterol, not just cardiovascular risk, is important in deciding who should receive statin treatment. *European Heart Journal*, *36*(43), 2975–2983.

Assessing net benefit of ASCVD risk estimation



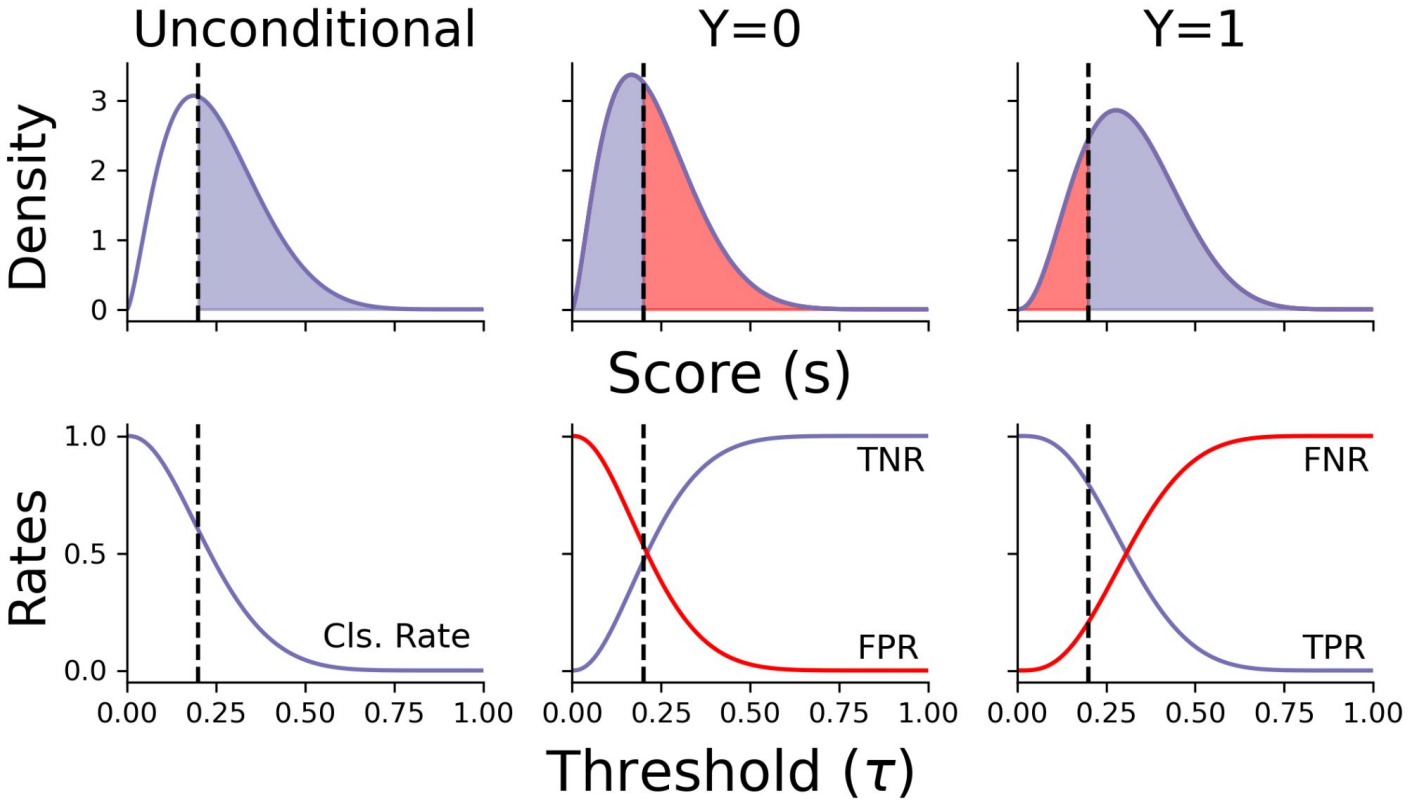
How should we interpret equalized odds?

Criteria	Definition	Interpretation
Metric parity	$g(\cdot) \perp A$	Metric g does not differ
Demographic parity	$S \perp A$	Score distribution does not differ
Demographic parity	$\hat{Y} \perp A$	Classification rate does not differ
Equalized odds	$S \perp A \mid Y$	Identical ROC curves
Equalized odds	$\hat{Y} \perp A \mid Y$	Equal TPR and equal FPR
Group calibration	$\mathbb{E}[Y \mid S = s, A] = s$	Calibrated for each group
Sufficiency	$Y \perp A \mid S$	Calibration curves do not differ
Predictive parity	$Y \perp A \mid \hat{Y} = 1$	PPV does not differ

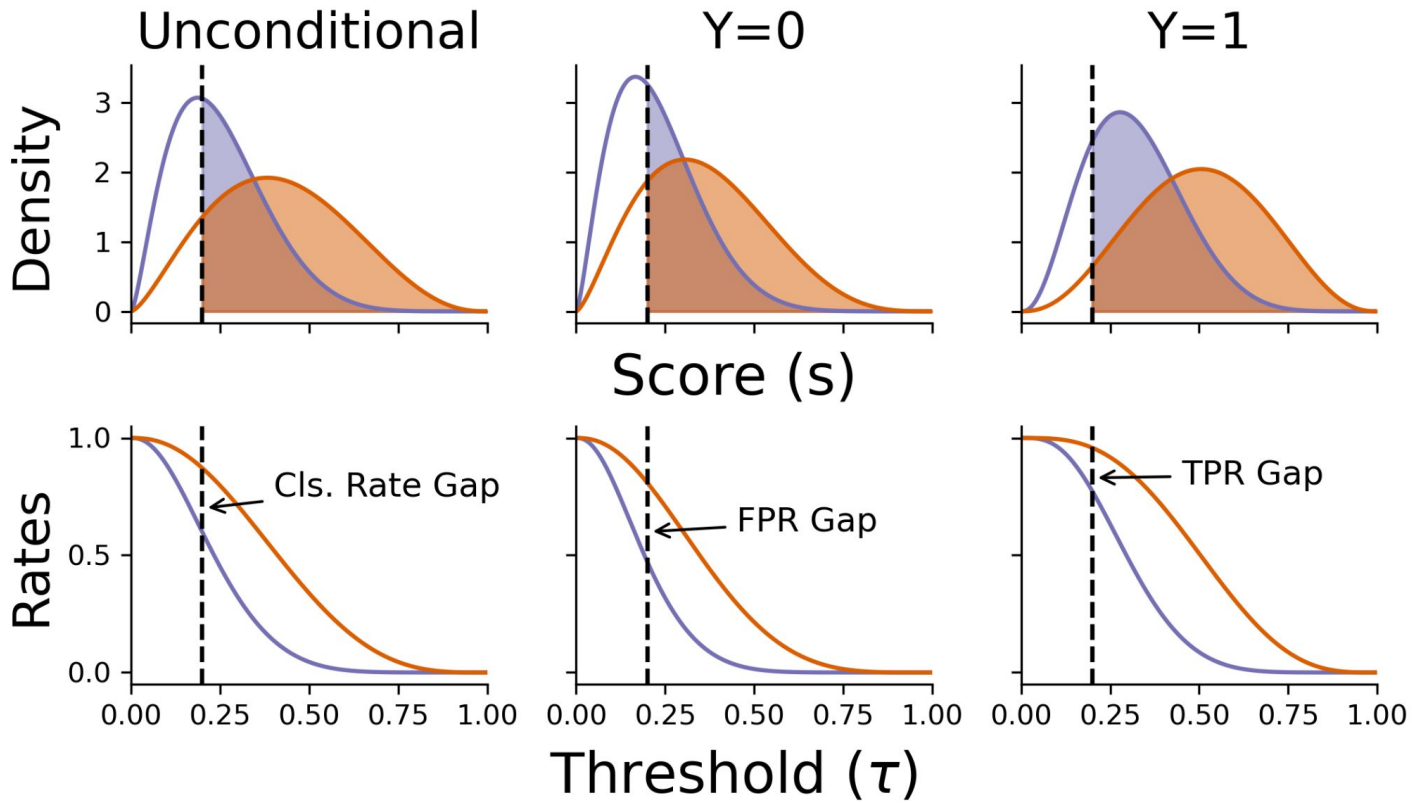
How should we interpret equalized odds violation?

- Violation is expected if incidence of the outcome differs and group calibration holds
- May or may not be problematic: requires assessment of *why* rates differ
  - Biases in the measurement of the outcome relative to the true construct of interest
  - Population-level differences in outcomes due to social determinants of health, structural racism, or other factors
  - Simple differences in model fit

Score distributions and conditional error rates

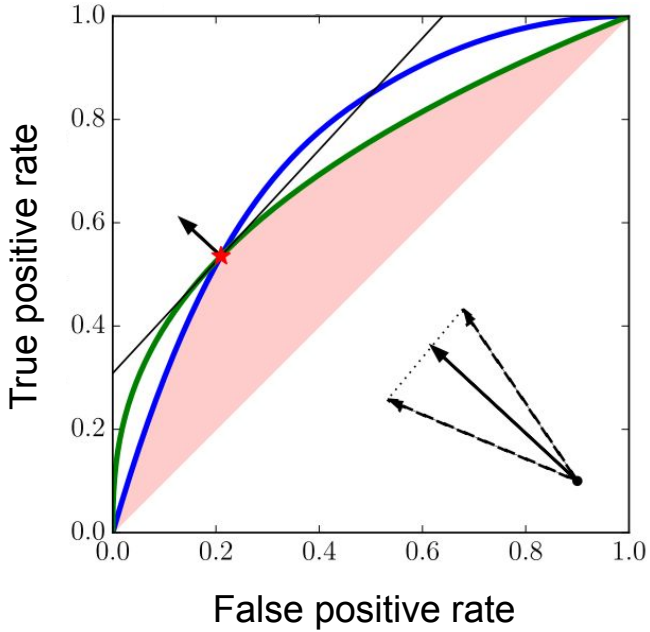


Fairness criteria over score distributions



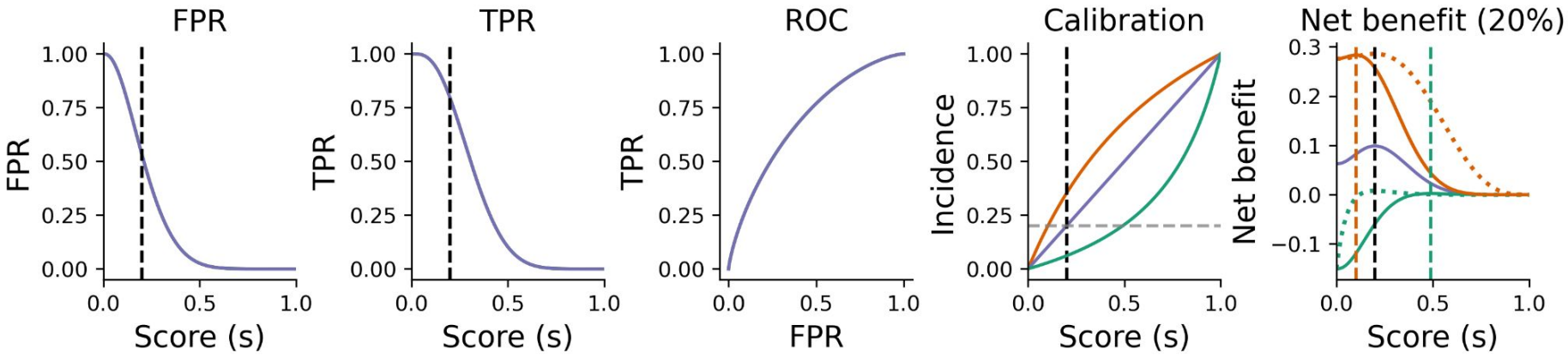
Is post-processing for equalized odds appropriate?

- Post-processing reduces net benefit relative to calibration and threshold-selection
  - Explicit threshold adjustment results in reduced net benefit
  - Transformations of the score result either in reduction in fit or implicit threshold adjustment through miscalibration



Hardt, M., Price, E., Srebro, N. N. N., & Others. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 3315–3323.

Miscalibration induced by post-processing for EO



Theoretical interpretation

- Empirical risk minimizer → model is calibrated and satisfies sufficiency
  - $\max(M_{\text{Cal}}, M_{\text{Suf}}) \leq O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*})$
- Empirical risk minimizer has non-trivial demographic parity and equalized odds violation when prevalence/incidence differs
  - $M_{\text{DemParity}} \geq k_{\text{rates}} - O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*})$
  - $M_{\text{EqualOdds}} \geq k_{\text{noise}} k_{\text{rates}} - O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*})$

Liu, L. T., Simchowitz, M., & Hardt, M. (2019). The Implicit Fairness Criterion of Unconstrained Learning. *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 4051–4060).

What about positive predictive value (PPV)?

Criteria	Definition	Interpretation
Metric parity	$g(\cdot) \perp A$	Metric g does not differ
Demographic parity	$S \perp A$	Score distribution does not differ
Demographic parity	$\hat{Y} \perp A$	Classification rate does not differ
Equalized odds	$S \perp A \mid Y$	Identical ROC curves
Equalized odds	$\hat{Y} \perp A \mid Y$	Equal TPR and equal FPR
Group calibration	$\mathbb{E}[Y \mid S = s, A] = s$	Calibrated for each group
Sufficiency	$Y \perp A \mid S$	Calibration curves do not differ
Predictive parity	$Y \perp A \mid \hat{Y} = 1$	PPV does not differ

Key points for evaluation with fairness criteria

- Calibration-based criteria are consistent with the use of the maximum utility decision rule given a model
- Calibration-based criteria can be misleading for poorly-fitting models
- Differences in TPR, FPR, PPV, or classification rate are expected if group calibration holds and outcome incidence differs
- Post-processing to satisfy equalized odds, demographic parity, or predictive parity typically reduce utility

What are the best practices for developing predictive models that enable fair clinical decision making?

- Should model training objectives include explicit fairness constraints?
- How can we build models that predict outcomes well for each group?

Key points for model development

1. Identify the set of models conferring the largest net benefit for each group
   a. Typically the best-fitting set of models that are calibrated for each group at relevant thresholds
2. Unpenalized ERM using entire dataset is typically sufficient, optionally recalibrating by group
3. Objectives that penalize equalized odds violation do not generally increase net benefit or model performance
4. Set thresholds based on calibration, intervention effectiveness, preferences, and cost, *not* targeted TPR, FPR, or PPV

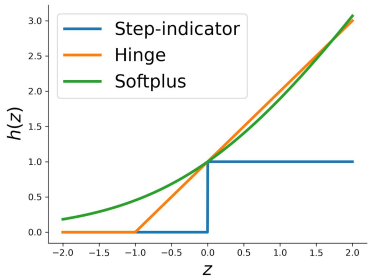
“In-processing” approaches for algorithmic fairness

- Constrained optimization
  - $\min_\theta \sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)),\ R < \kappa$
- Regularized learning objectives
  - $\min_\theta \sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) + \lambda R$
- Forms of ***R***
  - Divergence-based
    - $\frac{1}{K} \sum_{A_k \in \mathcal{A}} D(P(\cdot \mid A = A_k) \mid\mid P(\cdot))$
  - Metric-based
    - $\frac{1}{K} \sum_{A_k \in \mathcal{A}} \left(g(\mathcal{D}_k, f_\theta) - g(\mathcal{D}, f_\theta)\right)^2$

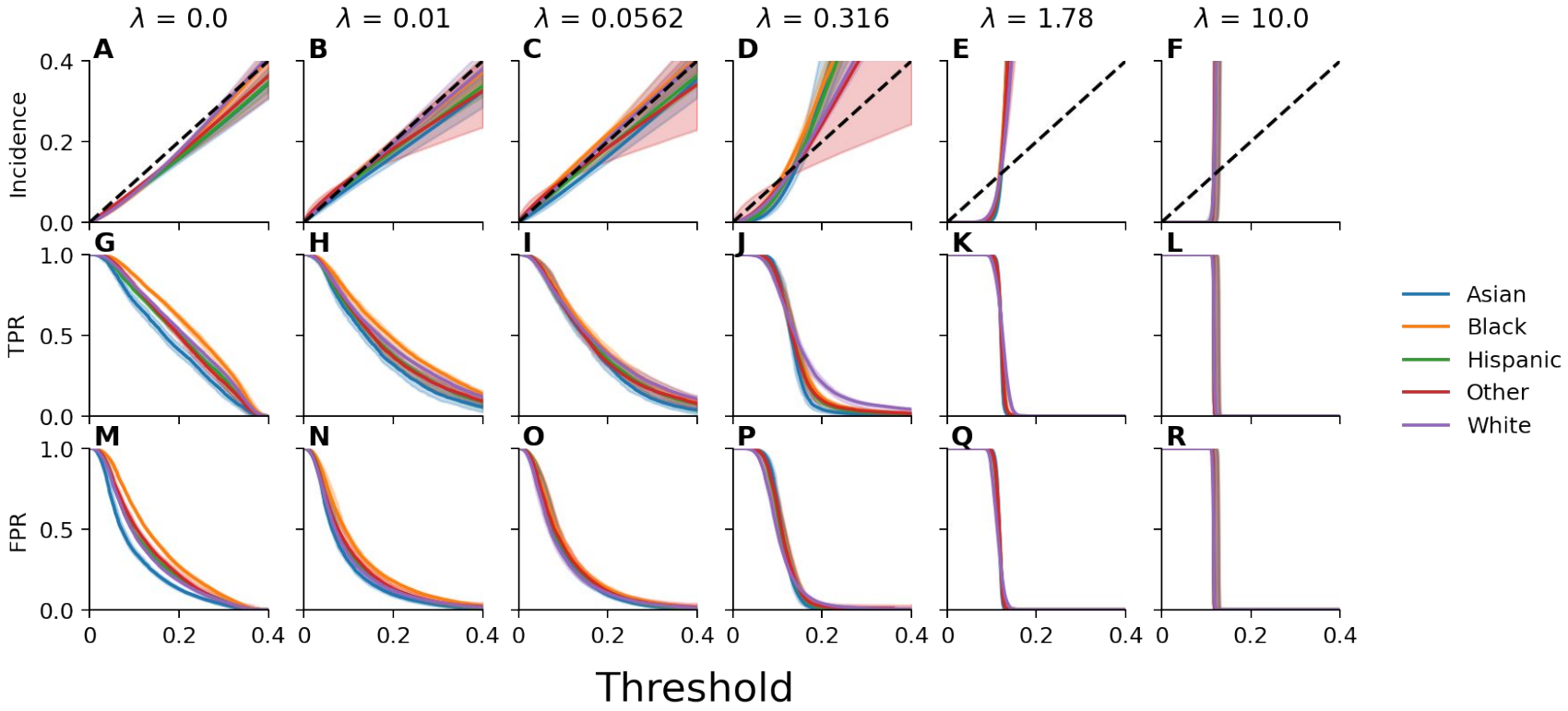
Pfohl, S. R., Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113, 103621.

Regularized learning objectives for fairness

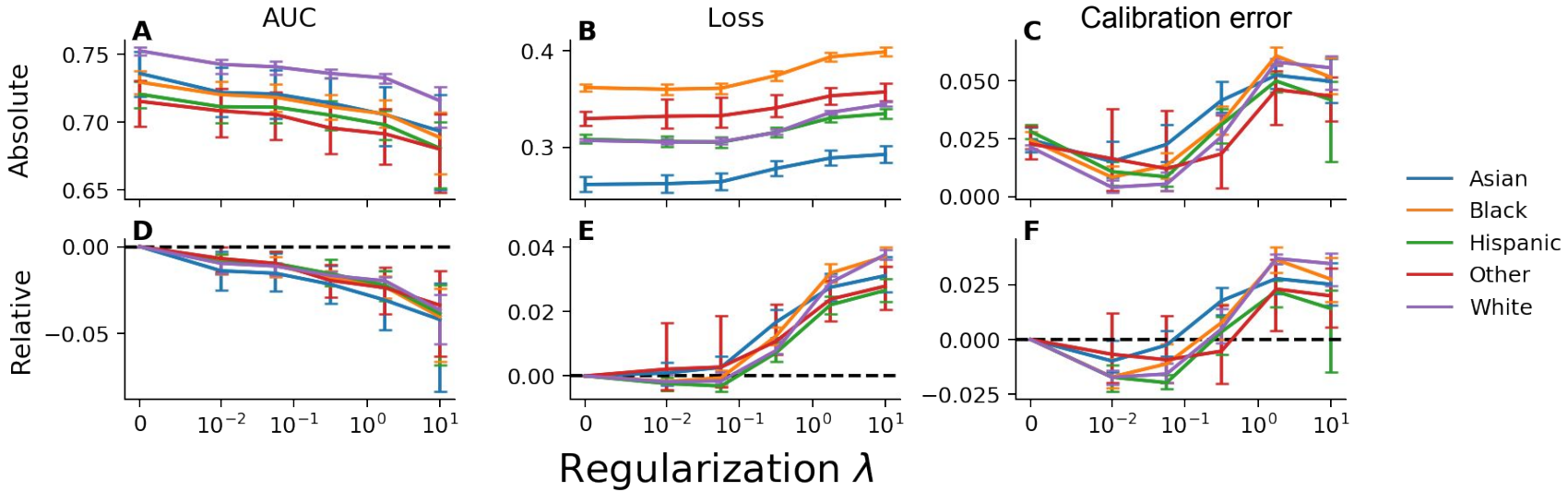
- For equalized odds (threshold-free)
  - Maximum mean discrepancy
    - $\frac{1}{K} \sum_{Y_j \in \mathcal{Y}} \sum_{A_k \in \mathcal{A}} \hat{D}_{\text{MMD}}(P(f(X) \mid A = A_k, Y = Y_j) \mid\mid P(f(X) \mid Y = Y_j))$
  - Difference in means
    - $\frac{1}{K} \sum_{Y_j \in \mathcal{Y}} \sum_{A_k \in \mathcal{A}} (\mathbb{E}_{X|A=A_k, Y=Y_j}[f_\theta(x)] - \mathbb{E}_{X|Y=Y_j}[f_\theta(x)])^2$
- Regularizers that incorporate rate-based metrics via differentiable surrogates
  - $g_{TPR} := \mathbb{E}_{x \sim \mathcal{D}|Y=1} h(f_\theta(x) - \tau)$
  - $g_{FPR} := \mathbb{E}_{x \sim \mathcal{D}|Y=0} h(f_\theta(x) - \tau)$
  - $g_{AUC} := \mathbb{E}_{x^1 \sim \mathcal{D}|Y=1} \mathbb{E}_{x^0 \sim \mathcal{D}|Y=0} h\left(f_\theta(x^1) - f_\theta(x^0)\right)$



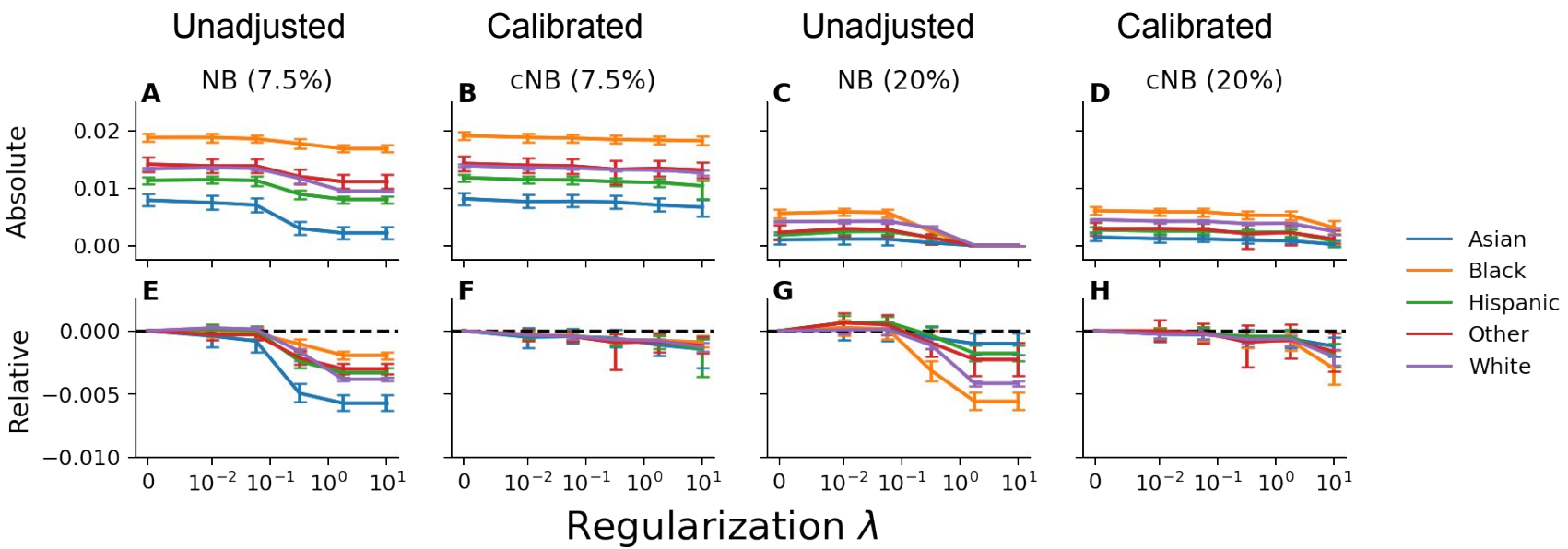
Effect of EO regularization on calibration, TPR, FPR



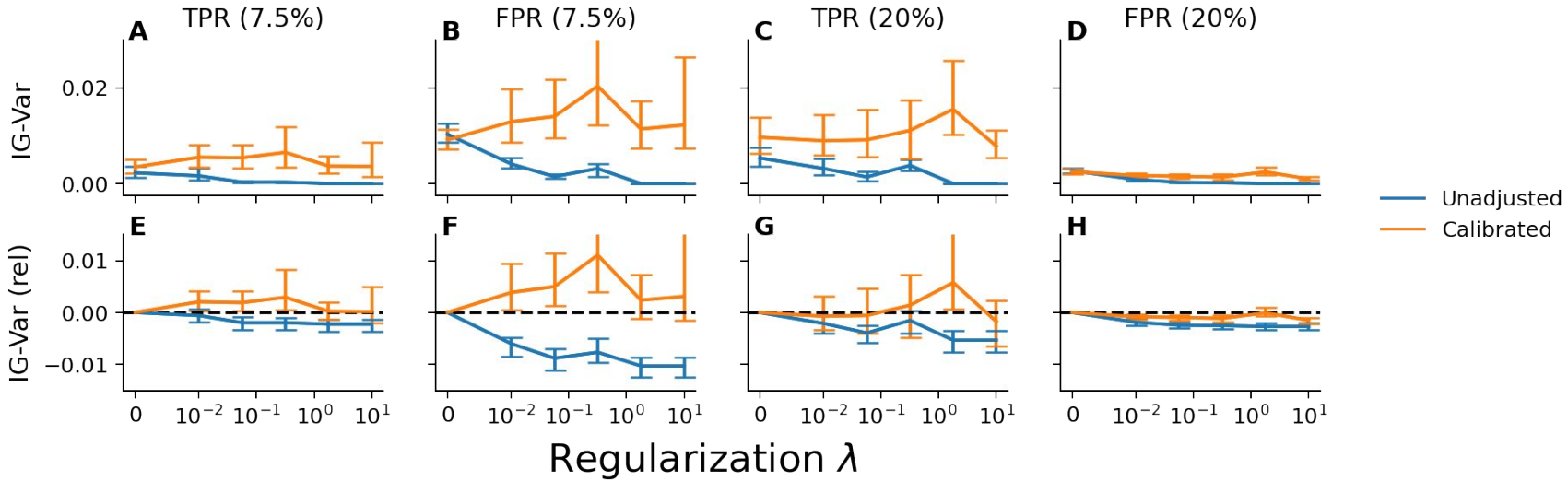
Effect of EO regularization on model performance



ERM confers the most *calibrated* net benefit



Recalibration does not preserve EO satisfaction



Should models include explicit fairness constraints?

- Equalized odds penalties
  - Typically confer less net benefit than unpenalized ERM after accounting for calibration
  - Any apparent benefits with minor amounts of regularization are due to miscalibration of the unpenalized ERM model
  - Strong penalties induce miscalibration and reduction in discriminatory capabilities of the model
- Calibration/sufficiency penalties or objectives
  - Not thoroughly tested or characterized, but important future direction
  - Value unclear due to consistency of calibration and sufficiency with loss minimization and the relative ease of calibration-based post-processing (e.g. recalibration or multi-calibration)

Approaches to target performance over groups

- Unconstrained empirical risk minimization (ERM) applied over the whole population (“pooled”)
- Unconstrained ERM applied separately for each group (“stratified”)
- Regularized learning objectives over model performance measures
  - Minimize differences in AUC and log-loss across groups
- Distributionally robust optimization (DRO) formulated to optimize to target worst-case performance over groups
  - Optimize worst-case AUC or log-loss across groups

Robustness to subpopulation shift

- Distributionally robust optimization for supervised learning
  - $\min_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}\, \ell(y, f_\theta(x))$
- Uncertainty set over shifts in the proportion of data from each group
  - $\mathcal{Q} := \{\sum_{k=1}^{K} \lambda_k P(X, Y \mid A = A_k) : \lambda \in \Lambda := \{\sum_{k=1}^{K} \lambda_k = 1; \lambda_k \geq 0\}\}$
- The *Group DRO* objective (Sagawa *et al,* 2020)
  - $\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \sum_{k=1}^{K} \lambda_k\, \mathbb{E}_{(x,y)\sim \mathcal{D}_k}\, \ell(y, f_\theta(x))$
- Implemented as alternating updates
  - Exponentiated gradient ascent over weights: $\lambda_k \leftarrow \lambda_k \exp(\eta \ell_k) / \sum_{k=1}^{K} \exp(\eta \ell_k)$
  - Weighted SGD: $\theta \leftarrow \theta - \eta \nabla_\theta \sum_{k=1}^{K} \lambda_k \ell_k$

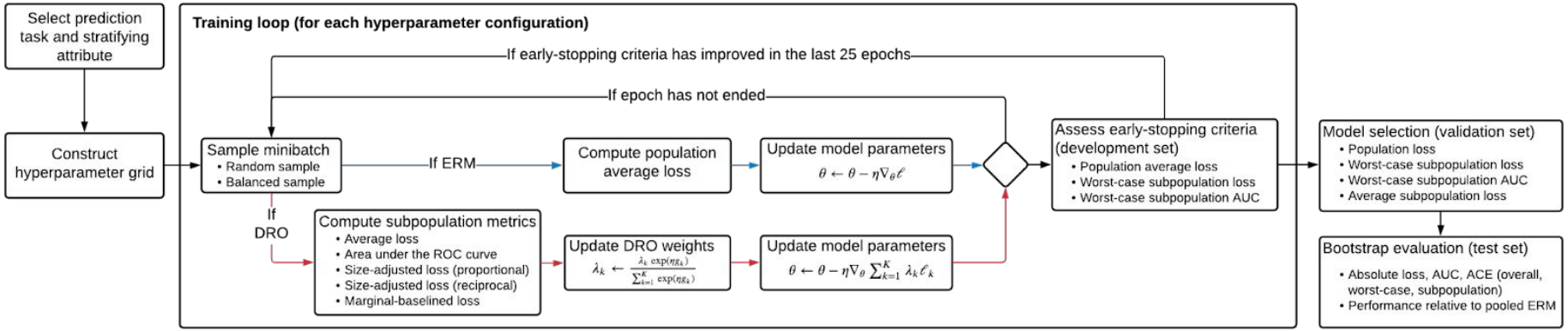
Sagawa, Shiori, et al. “Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization” *International Conference on Learning Representations*. 2020.

Flexible DRO objectives

- General form of the update
  - Update over lambda: $\lambda_k \leftarrow \lambda_k \exp(\eta g(\mathcal{D}_k, f_\theta)) / \sum_{k=1}^{K} \exp(\eta g(\mathcal{D}_k, f_\theta))$
  - Update over model parameters (unchanged): $\theta \leftarrow \theta - \eta \nabla_\theta \sum_{k=1}^{K} \lambda_k \ell_k$
- An AUC-based objective (no surrogates)
  - $g_{\text{AUC}}(\mathcal{D}_{A_k}, f_\theta) = 1 - \frac{1}{n_k^{y=1} n_k^{y=0}} \sum_{i=1}^{n_k^{y=1}} \sum_{j=1}^{n_k^{y=0}} \mathbb{I}(f_\theta(x_i) > f_\theta(x_j))$

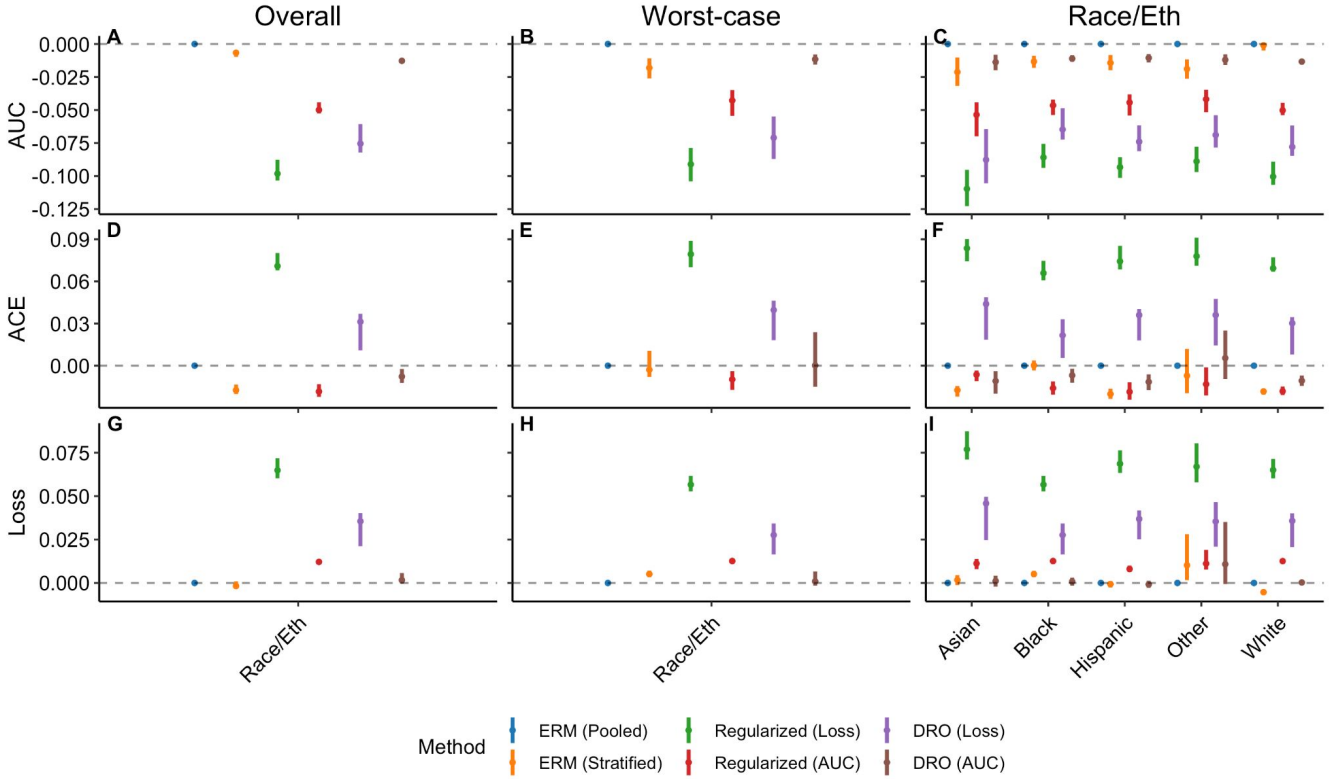
Pfohl, S. R., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., & Shah, N. H. (2021). A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *arXiv preprint arXiv:2108.12250*.

Model training and evaluation workflow



Pfohl, S. R., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., & Shah, N. H. (2021). A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *arXiv preprint arXiv:2108.12250*.

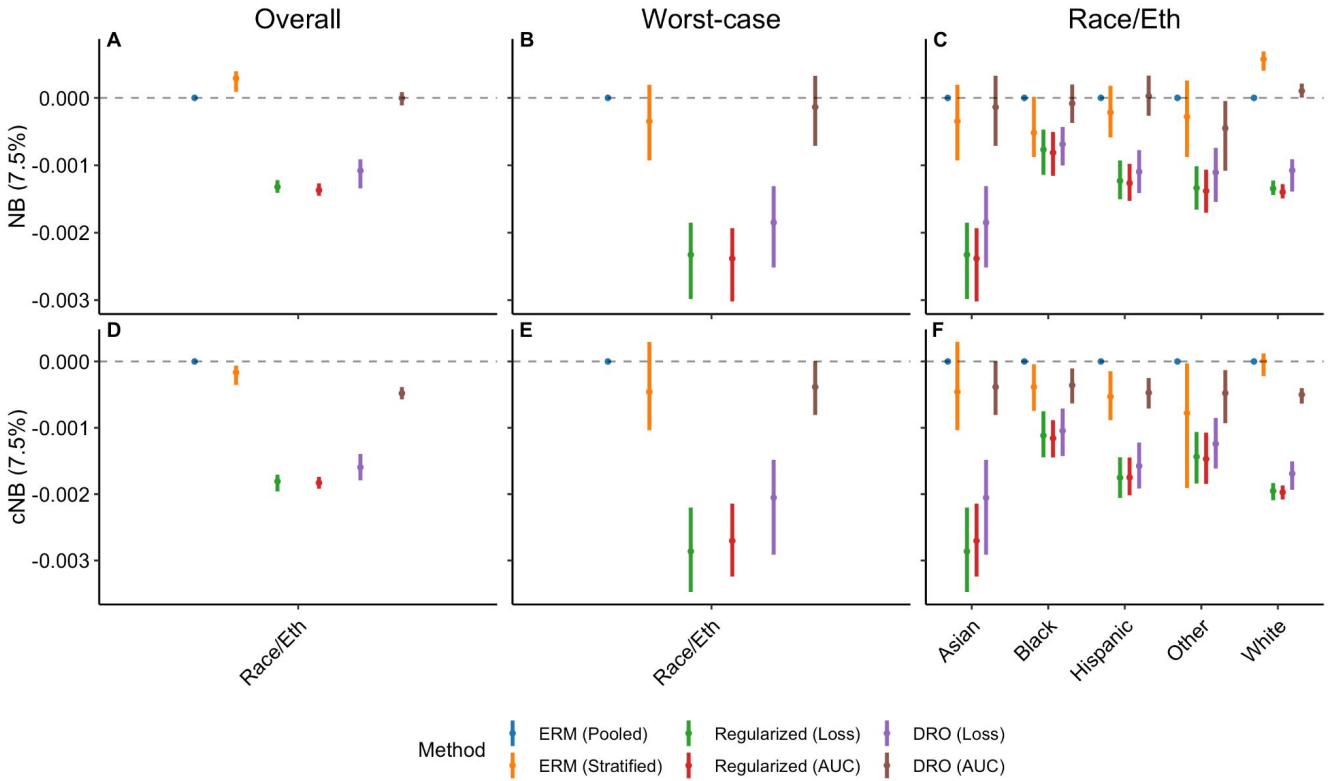
Comparison of approaches - performance



Comparison of approaches - net benefit

Unadjusted

Calibrated



Can we improve on standard paradigms?

- Approaches that target differences or worst-case performance measures do not generally confer more net benefit than ERM after accounting for miscalibration
- In practice, DRO approaches that target worst-case performance do not seem to improve worst-case performance over simpler ERM strategies
  - Does not necessarily preclude further methods development
- Calibration can be improved over ERM in some instances
  - Stratified training
  - AUC-based penalties and DRO objectives

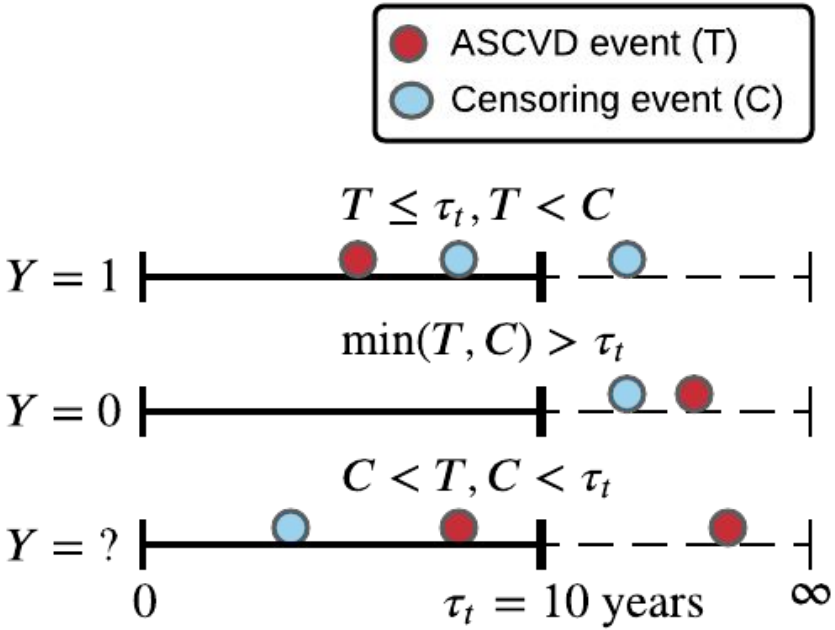
Summary of key takeaways and recommendations

1. Transparently document protocols and problem formulation
2. Learn the best-fitting set of models for each population that are calibrated at relevant thresholds
3. Comprehensively report and contextualize stratified performance metrics
4. Prioritize calibration-based fairness assessments
5. Do not consider context-free fairness assessments as sole indicators of whether ML-intervention introduces/exacerbates harm or is equity-promoting
6. Current approaches to algorithmic fairness are limited in scope, but can still be useful

Additional content

Accounting for censoring in ten-year ASCVD

- Ten-year ASCVD outcomes are often not fully observed due to censoring
- Can represent as a binary outcome if censoring is accounted for
- Use inverse probability of censoring weights (IPCW) for all training objectives and evaluation metrics
- Training objective: weighted empirical risk minimization (ERM)
  - $\min_{\theta \in \Theta} \sum_{i=1}^{N} w_i \ell(y_i, f_\theta(x_i))$



Inverse probability of censoring weighting (IPCW)

- Weights scale inversely with the probability of remaining uncensored
- Weighted empirical risk minimization (ERM)
  - $\min_{\theta \in \Theta} \sum_{i=1}^{N} w_i \ell(y_i, f_\theta(x_i))$
- An IPCW-weighted variant of each training objective and evaluation metric is used

Definitions

$$u_i^y = \min(t_i, c_i, \tau_t)$$

$$G(u, x) \approx P(C > u \mid X = x)$$

$$\delta_i^y = 1 - \mathbb{I}[c_i < t_i] * \mathbb{I}[c_i < \tau_t]$$

$$w_i = \frac{\delta_i^y}{G(u_i^y, x_i)} \left(\sum_{i=1}^{N} \frac{\delta_i^y}{G(u_i^y, x_i)} \right)^{-1}$$

Assume

$$T \perp C \mid X$$

$$\delta_i^y = 1 \rightarrow G(u_i^y, x_i) > 0$$

Aggregate expected utility and net benefit

- Aggregate utility is the expected utility of the decision rule over a population

$$U_{\text{agg}}(\tau) = \mathbb{E}[U^1_{\text{cond}} \mid S \geq \tau]P(S \geq \tau) + \mathbb{E}[U^0_{\text{cond}} \mid S < \tau]P(S < \tau)$$

- Net benefit is a normalized aggregate utility metric that parametrizes a utility function based on an assumed optimal threshold

$$\text{NB}(\tau; \tau^*) = P(S \geq \tau \mid Y = 1)P(Y = 1) - P(S \geq \tau \mid Y = 0)P(Y = 0)\frac{\tau^*}{1-\tau^*}$$

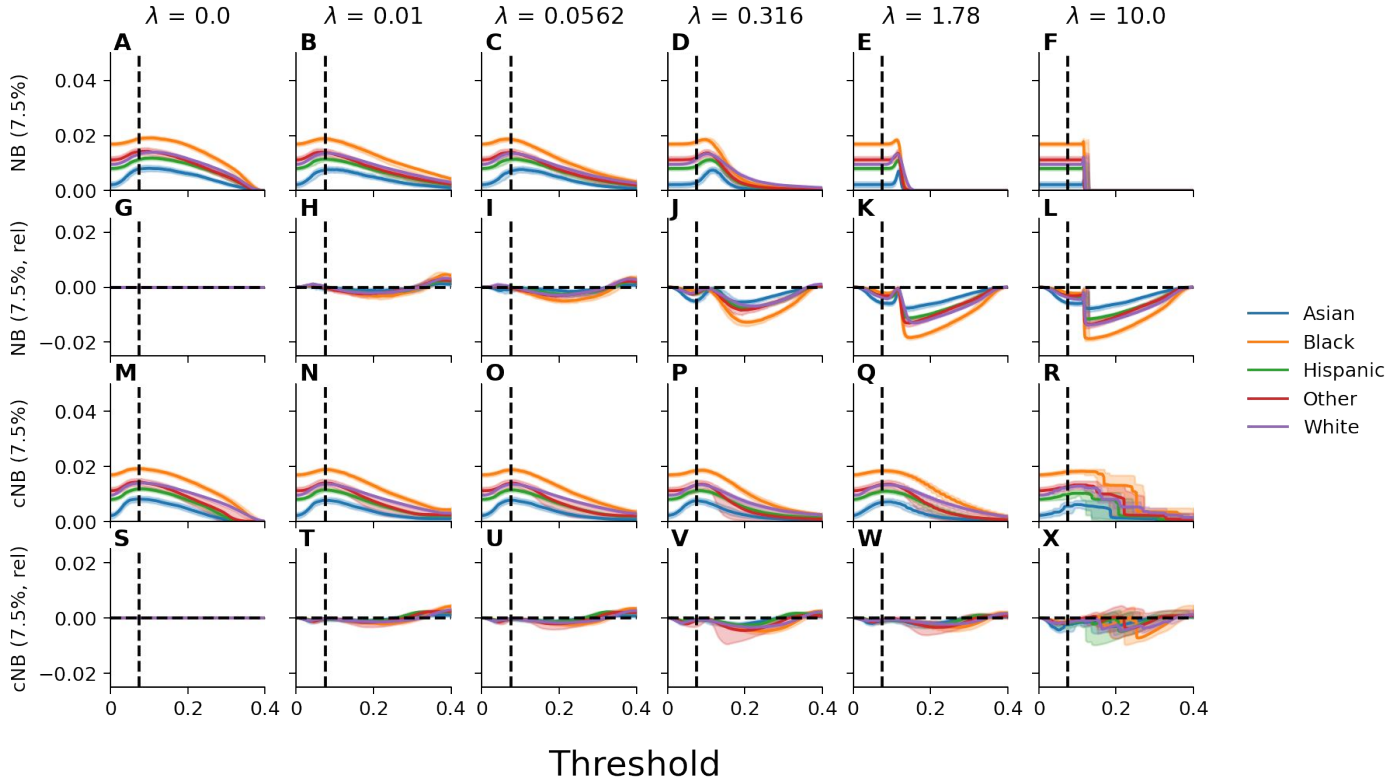
Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." Medical Decision Making 26.6 (2006): 565-574.

Net benefit of statin initiation

- Net benefit assuming expected harm is independent of risk estimate
  - $\mathrm{NB}(\tau_y;\tau_y^*) = -\mathbb{E}[p_y^0(s) \mid S < \tau_y]P(S < \tau_y) - \mathbb{E}[p_y^1(s) \mid S \geq \tau_y]P(S \geq \tau_y) - \mathrm{ARR}(\tau_y^*)P(S \geq \tau_y) + P(Y = 1)$
- If relative risk reduction is constant
  - $\mathrm{NB}(\tau_y;\tau_y^*) = -(1 - \mathrm{NPV}(\tau_y))P(S < \tau_y) - P(S \geq \tau_y)\Big((1-r)\mathrm{PPV}(\tau_y) + r\tau_y^*\Big) + P(Y = 1)$
- Simplifying assumptions, following Soran *et. al* [1]
  - Moderate intensity statin (20 mg atorvastatin) → 43% reduction in LDL-C
  - Each 1 mmol/L reduction in LDL-C → 22% relative reduction in ten-year risk
  - Approximate LDL-C ~ predicted risk as a constant and compute population mean of most recent LDL-C measurement at prediction time → 3.01 mmol/L
  - Compute relative risk reduction: $r = 1 - (1 - 0.22)^{(3.01 * 0.43)} = 0.275 = 27.5\%$

[1] Soran, H., Schofield, J. D., & Durrington, P. N. (2015). Cholesterol, not just cardiovascular risk, is important in deciding who should receive statin treatment. *European Heart Journal*, *36*(43), 2975–2983.

Detailed view of effect on net benefit at 7.5%

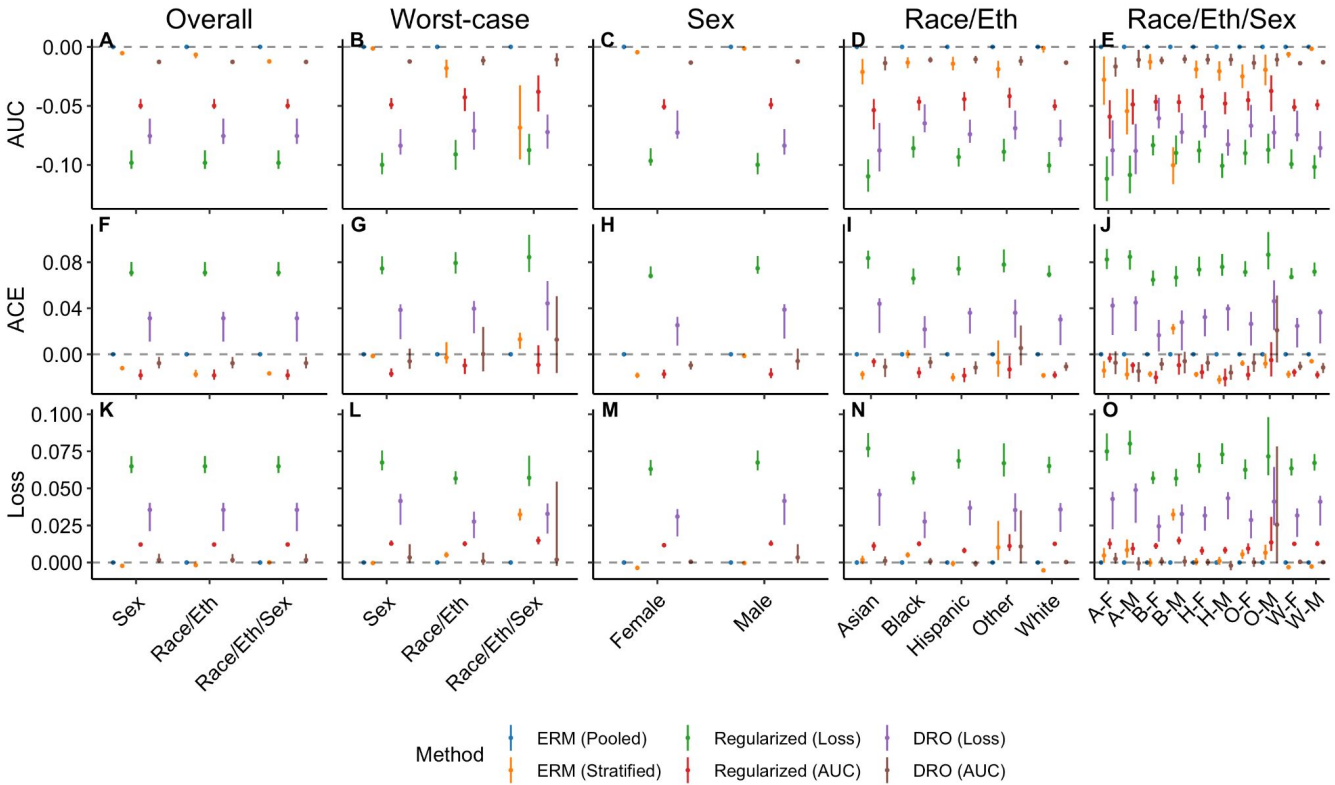


DRO over groups with censored binary outcomes

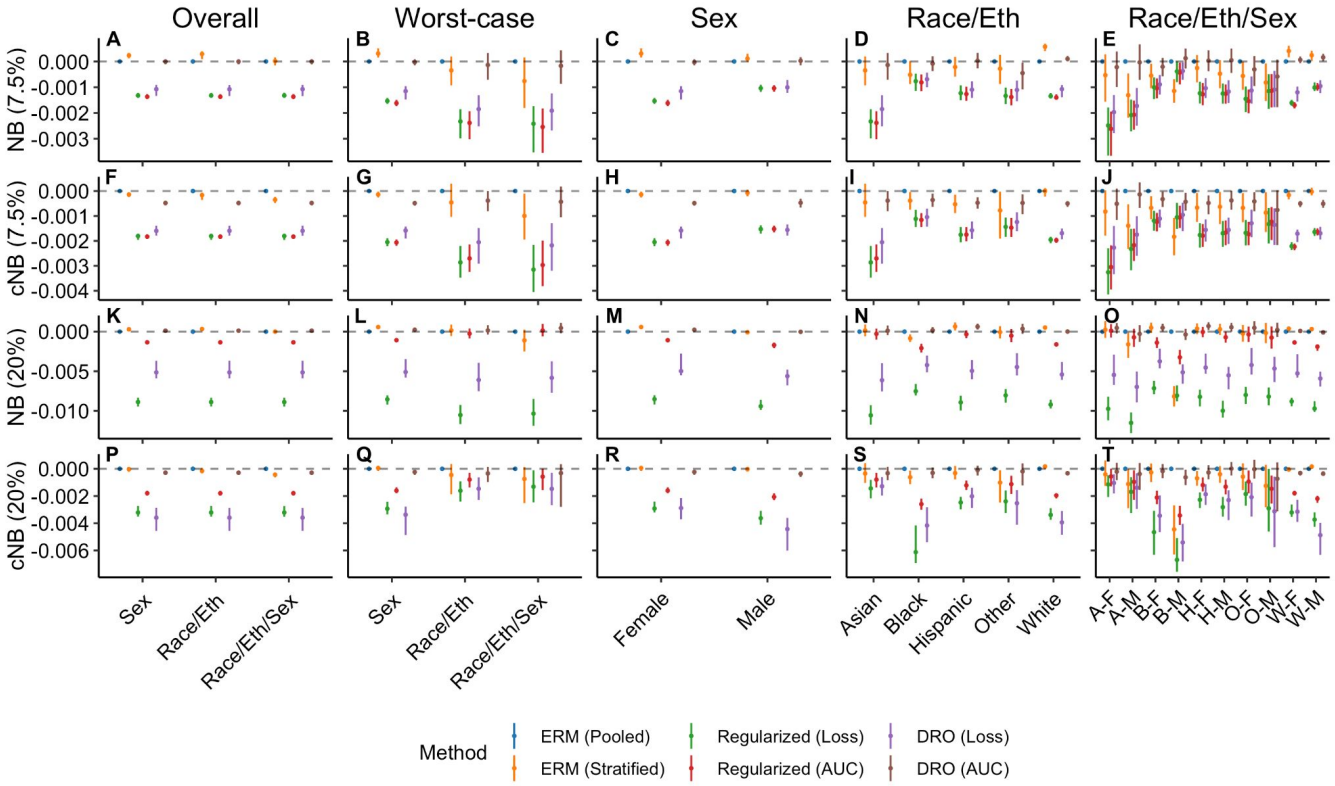
- IPCW variants of the alternating updates
  - $\lambda_k \leftarrow \lambda_k \exp\left(\eta \sum_{i=1}^{n_k} w_i \ell\left(y_i, f_\theta(x_i)\right)\right) \Big/ \sum_{k=1}^{K} \exp\left(\eta \sum_{i=1}^{n_k} w_i \ell\left(y_i, f_\theta(x_i)\right)\right)$
  - Update on model parameters: $\min_{\theta \in \Theta} \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n_k} w_i \ell(y_i, f_\theta(x_i))$

- Flexible DRO objective with IPCW-weighted AUC
  - $g_{\text{AUC}}^{\text{IPCW}}(\mathcal{D}_{A_k}, f_\theta) = \sum_{i=1}^{n_k^{y=1}} \sum_{j=1}^{n_k^{y=0}} w_{ij} \mathbb{I}(f_\theta(x_i) > f_\theta(x_j))$
  - $w_{ij} = \frac{\delta_i^y}{G(u_i^y, x_i)} \frac{\delta_j^y}{G(u_j^y, x_j)} \left(\sum_{i=1}^{n_k^{y=1}} \sum_{j=1}^{n_k^{y=0}} \frac{\delta_i^y}{G(u_i^y, x_i)} \frac{\delta_j^y}{G(u_j^y, x_j)} \right)^{-1}$

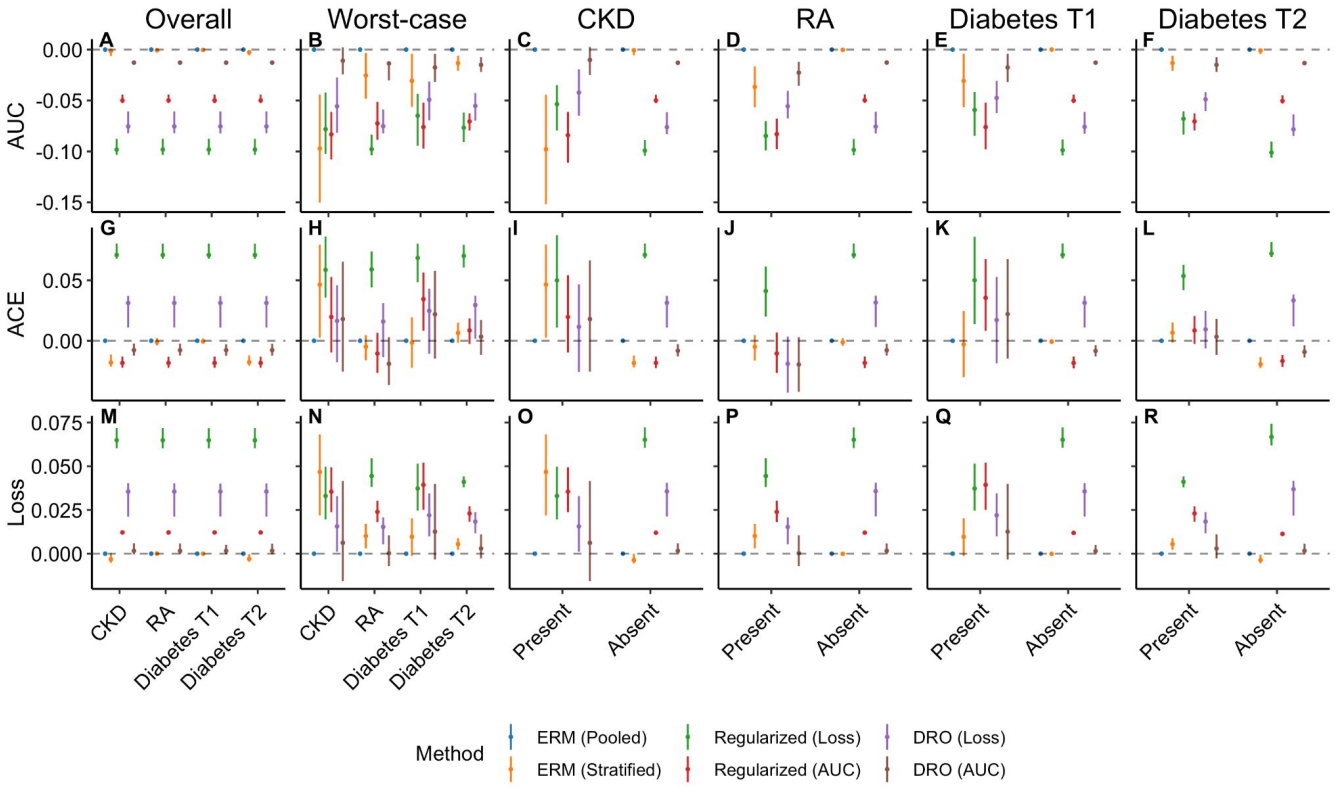
Comparison of approaches - performance



Comparison of approaches - net benefit



Comparison of approaches - comorbidities



Comparison of approaches - comorbidities