Is this Model Reliable for Everyone? Testing for Strong Calibration

Jean Feng[1] **Alexej Gossmann**[2] **Romain Pirracchio**[1]
Nicholas Petrick[2] **Gene Pennello**[2] **Berkman Sahiner**[2]
[1]University of California, San Francisco
[2]U.S. Food and Drug Administration, Center for Devices and Radiological Health

Abstract

In a well-calibrated risk prediction model, the average predicted probability is close to the true event rate for any given subgroup. Such models are reliable across heterogeneous populations and satisfy strong notions of algorithmic fairness. However, the task of auditing a model for strong calibration is well-known to be difficult—particularly for machine learning (ML) algorithms—due to the sheer number of potential subgroups. As such, common practice is to only assess calibration with respect to a few predefined subgroups. Recent developments in goodness-of-fit testing offer potential solutions but are not designed for settings with weak signal or where the poorly calibrated subgroup is small, as they either overly subdivide the data or fail to divide the data at all. We introduce a new testing procedure based on the following insight: if we can reorder observations by their expected residuals, there should be a change in the association between the predicted and observed residuals along this sequence if a poorly calibrated subgroup exists. This lets us reframe the problem of calibration testing into one of change-point detection, for which powerful methods already exist. We begin with introducing a sample-splitting procedure where a portion of the data is used to train a suite of candidate models for predicting the residual, and the remaining data are used to perform a score-based cumulative sum (CUSUM) test. To further improve power, we then extend this adaptive CUSUM test to incorporate cross-validation, while maintaining Type I error control under minimal assumptions. Compared to existing methods, the proposed procedure consistently achieved higher power in empirical analyses.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Calibration is a fundamental measure of model reliability: a risk prediction model is called "calibrated" or "reliable" for a subgroup if the average predicted probability corresponds to the observed event rate [Foster and Vohra, 1998]. When decisions are made using absolute risk thresholds—as is common in medicine [Goff et al., 2014]—calibration directly impacts the utility of a model [Van Calster and Vickers, 2015]. However, machine learning (ML) algorithms are typically trained to optimize average performance and can be poorly calibrated in particular subgroups [Chatterjee et al., 2016, Barda et al., 2021], leading to concerns regarding their robustness and fairness. In fact, performance can be particularly low for subgroups defined by interactions of multiple variables (e.g. race and gender), an issue known as intersectionality [Buolamwini and Gebru, 2018]. Ideally, a risk prediction model is "strongly" calibrated, in that it is calibrated for all individuals or, equivalently, all possible subgroups [Van Calster et al., 2016, Zhao et al., 2020].

Unfortunately, achieving or even verifying strong calibration is challenging due to the curse of dimensionality: as the number of variables grows, the subgroups—and thus the number of observations per subgroup—get smaller. As such, much of previous research has been focused on "moderate" calibration, where the subgroups are defined as observations with similar risk predictions [Cox, 1958, Brown et al., 1975, Tsiatis, 1980, Hawkins, 1991, Hosmer et al., 1997, Hosmer and Hjort, 2002, Lin et al., 2002, Widmann et al., 2019, DiCiccio et al., 2020, Hudson et al., 2021, Lee et al.,

2022, Glaser et al., 2023].[1] These methods are widely used among practitioners, given their ease of use and applicability to small datasets. With the recent interest in algorithmic fairness and model reliability, many recent works have sought to achieve stronger forms of calibration, either by identifying subgroups that require revision [Chung et al., 2019, Eyuboglu et al., 2022] or revising the model directly [Hebert-Johnson et al., 2018, Kim et al., 2019, Luo et al., 2022]. However, these methods are either meant to be exploratory or only provide statistical guarantees when the number of observations is sufficiently large. The minimum sample size is typically at least tens or even hundreds of thousands of observations, which is unrealistic in many settings.

Rather than tackling the difficult task of subgroup identification or model revision, we instead consider the problem of testing if there *exists* any poorly calibrated subgroup. This is much more feasible in settings with limited data or lower signal-to-noise ratios, and still answers the important yes/no question "Is this ML algorithm reliable for everyone?" Moreover, the answer to this question can help decide if more sophisticated data-hungry procedures are necessary.

We formalize this as the following hypothesis test. For binary classification tasks, we represent a random observation from the population by a random vector $X \in \mathbb{R}^d$ and binary random variable Y. Let $\hat{p} : \mathcal{X} \mapsto [0,1]$ be the risk prediction algorithm and $p_0 : \mathcal{X} \mapsto [0,1]$ be the true event rate (i.e. $p_0(x) = \Pr(Y = 1|X = x)$) over some domain $\mathcal{X} \subseteq \mathbb{R}^d$. For some pre-specified tolerance level $\delta \geq 0$, define the poorly calibrated subgroup as $A_\delta = \{x \in \mathcal{X} : |\hat{p}(x) - p_0(x)| > \delta\}$. The hypothesis test checks if the set A_δ is too large, i.e.

$$\begin{aligned} H_0 &: \Pr(X \in A_\delta) \leq \epsilon \\ H_1 &: \Pr(X \in A_\delta) > \epsilon \end{aligned} \tag{1}$$

from some minimally acceptable prevalence $\epsilon \geq 0$. Prior works have primarily tested the special case where $d = 1$ and $\epsilon = \delta = 0$ [Hudson et al., 2021]. For instance, tests for moderate calibration often consider the strict null hypothesis $H_0 : \mathbb{E}[Y = 1|\hat{p}(x) = q] = q$ for all $q \in [0,1]$, which corresponds to representing each observation by its predicted probability. However, the strict null can be rejected even if model miscalibration is not practically significant, so interpretation of such tests can be challenging. It is not clear how to extend these works to settings with $\delta > 0$ or for much larger d.

Recent advancements in the goodness-of-fit (GOF) testing literature provide potential solutions [Janková et al., 2020, Zhang et al., 2021]. Although GOF tests technically answer a different question from (1), one could consider extending the methods proposed in these works to test for strong calibration. The main idea in these works is to break the curse of dimensionality using sample-splitting: they use one partition of the data to train a model of the residuals and the remaining partition to test for GOF with respect to the learned residual model. The difference between [Janková et al., 2020] and [Zhang et al., 2021] is primarily in the second step. The former assesses the association between the predicted and observed residuals with respect to the entire population through a score test. The latter bins observations with similar predicted residuals and performs a Chi-squared test.

In settings where only a small subgroup is miscalibrated, it is critical that we reduce any wastage of power. The aforementioned GOF tests were not designed for this setting, so their power for detecting miscalibration in subgroups is limited. [Janková et al., 2020] calculates the average score, which overlooks variation across subgroups. [Zhang et al., 2021] bins observations into subgroups, but no information is borrowed across bins and the procedure is highly sensitive to the number of bins; the procedure may accidentally divide the poorly calibrated subgroup, thereby reducing its power to detect miscalibration. In addition, these works only use random forests (RFs) to model the residuals but, as we show later, RFs are weak at extracting the remaining signal after a tree-based model is fitted to the data. Instead, it is important to consider a *diverse* pool of candidate residual models. Also, these tests only perform a single sample-split. To extend these procedures to use cross-validation (CV), one must account for the correlation between residual models trained across folds. Finally, to tune hyperparameters of the residual model, both procedures perform further splitting of the training data. This is not only noisy in small sample sizes but also computationally expensive when combined with CV.

We introduce a more powerful testing procedure for strong calibration motivated with the following insight: if observations are ordered by their predicted residuals, we expect the association between the observed and predicted residuals to drop somewhere along this sequence if a poorly calibrated subgroup exists. Our key contributions are (i) we show how reframing the problem of detecting a poorly calibrated subgroup into that of changepoint detection substantially improves power because the changepoint structure closely mimics the true subpopulation structure; (ii) we demonstrate how additional gains in power and computational efficiency can be made by fitting

[1] Here we follow the calibration hierarchy defined in [Van Calster and Vickers, 2015]. Note that some works refer to "moderate calibration" in this hierarchy as "strong calibration."

a pool of candidate residual models and performing a suite of structural change tests; (iii) we incorporate CV to further improve power, while maintaining Type I error control under much weaker assumptions than prior works; and (iv) we provide visualization tools to aid model diagnosis. In experiments, the proposed procedure significantly outperforms existing methods. Code for reproducing all experiments is available at `https://github.com/jjfeng/testing_strong_calibration`.

1.1 Other related works

In the Introduction, we already discussed how this work relates to prior works on testing model calibration. Here we discuss some other related areas.

Multicalibration. Given an ML algorithm, multicalibration methods introduce post-hoc updates to achieve calibration across a rich collection of subgroups, typically using a boosting-type procedure [Hebert-Johnson et al., 2018, Kim et al., 2019, Gopalan et al., 2023, Globus-Harris et al., 2023]. To prevent overfitting, these methods typically leverage differential privacy methods and/or split the data into many partitions. As such, these methods require immense sample sizes [Barda et al., 2021, Kim et al., 2022]. In contrast, this paper is concerned with testing for the existence of a poorly calibrated subgroup, which is much more feasible in settings with limited data or lower signal-to-noise ratios.

Individual/metric fairness. Strong calibration of an ML algorithm can be viewed as a generalization of predictive parity for binary classifiers [Mitchell et al., 2021], which is only one approach to measuring model fairness. Other common measures of algorithmic fairness are concerned with statistical parity or balance of error rates between subgroups [Hardt et al., 2016, Mitchell et al., 2021]. Similar to the critiques of moderate calibration, group-wise equality in error rates has been critized for being too coarse [Dwork et al., 2012]. Recent works aim for individual or metric fairness to ensure similar performance between similar individuals [Ilvento, 2020, Ruoss et al., 2020] and respective hypothesis tests have been developed [Xue et al., 2020, Maity et al., 2021]. However, unlike the proposed procedure, these methods assume a similarity metric is known a priori, which corresponds to prespecifying the subgroup structure.

Conformal inference. Recent works have highlighted how the coverage rate guarantees from conformal inference procedures can be used to calibrate risk prediction algorithms [Vovk et al., 2020, Marx et al., 2022]. Ordinary conformal inference procedures only guarantee marginal coverage rates [Vovk et al., 2005], which satisfy notions of weak calibration [Van Calster et al., 2016]. More recent works have extended these methods to provide guarantees with respect to predefined subgroups [Vovk, 2013, Lei and Wasserman, 2014, Romano et al., 2020] and weighted neighborhoods [Guan, 2023]. Taking such guarantees to the limit, [Foygel Barber et al., 2021] proved that it is impossible for a non-trivial procedure to guarantee uniform conditional coverage rates. Our ability to test for strong calibration does not contradict this impossibility result and provides instead a complementary (and perhaps more positive) result. Because hypothesis tests start from the angle of "innocent until proven guilty," we can at least determine if there is sufficient evidence that a given model fails to satisfy strong calibration.

Distributionally robust optimization (DRO). DRO methods aim to train models that minimize the worst-case performance over some set of distributional perturbations [Ben-Tal et al., 2013, Duchi and Namkoong, 2021, Duchi et al., 2022]. Based on these ideas, recent works propose to estimate the worst-case performance of a given ML algorithm over all subgroups with size $\epsilon > 0$ and even provide confidence intervals [Subbaswamy et al., 2021, Li et al., 2021]. Nevertheless, to achieve valid statistical inference, these methods require much larger sample sizes, ϵ to be bounded away from zero, and the error model to converge at a fast enough rate. In contrast, our proposed procedure is suitable for smaller sample sizes, can test for arbitrarily small subgroups, and provides Type I error control under much weaker assumptions.

2 METHOD

For ease of exposition, we begin with the one-sided testing problem where we replace A_δ with the one-sided violation set $A_{\delta,>} = \{x \in \mathcal{X} : p_0(x) - \hat{p}(x) > \delta\}$. The first step is to reformulate the hypothesis test as a score test. In the main text of this paper, we focus on the test where $\epsilon = 0$. In the Appendix, we describe how the proposed procedure can be easily extended to address non-zero ϵ.

Let $\mathcal{H}_+$ be the class of bounded non-negative real-valued functions. For a given $h \in \mathcal{H}_+$, define a working model for the structural change of the log odds (logit) to be

$$\text{logit}\left(p(X;h)\right) := \text{logit}(\hat{p}_\delta(X)) + \theta h(X), \qquad (2)$$

where $\hat{p}_\delta(X) = [\hat{p}(X) + \delta]_{[0,1]}$ and $q \mapsto [q]_{[0,1]}$ is a projection into the range of valid probabilities $[0, 1]$. The gradient of the log likelihood, also known as the

score, at $\theta = 0$ is equal to

$$\dot{\ell}(Y|X;h) = \frac{\partial}{\partial\theta} \log p(Y|X;h)\Big|_{\theta=0} = (Y - \hat{p}_\delta(X))\, h(X).$$

In the set $A_{\delta,>}$, the expected score $\mathbb{E}[\dot{\ell}(Y|X;h)|X]$ is positive if $h(X)$ is positive. Outside of this set, the expected score is non-positive. As such, we will refer to h as a detector. We can rewrite the one-sided hypothesis test in terms of the maximum expected score over detectors in $\mathcal{H}_+$, i.e.

$$\begin{aligned} H_{0,>} &: \sup_{h\in\mathcal{H}_+} \mathbb{E}\left[(Y - \hat{p}_\delta(X))\, h(X)\right] \le 0 \\ H_{1,>} &: \sup_{h\in\mathcal{H}_+} \mathbb{E}\left[(Y - \hat{p}_\delta(X))\, h(X)\right] > 0. \end{aligned} \tag{3}$$

In practice, it is computationally infeasible to test the entire set $\mathcal{H}_+$. Instead, we will generate a subset $\widehat{\mathcal{H}}_+ \subseteq \mathcal{H}_+$ to replace $\mathcal{H}_+$ in (3), resulting in a *restricted* score test. Thus a procedure with Type I error control for a restricted score test also satisfies Type I error control for (3).

In the following sections, we introduce the testing procedure using single sample-split, extend it to incorporate CV, and finally extend it to the two-sided setting.

2.1 Sample-splitting

Suppose the audit data are composed of independent and identically distributed (IID) observations with variables $X_i \in \mathcal{X}$ and binary outcome Y_i for $i = 1, \cdots, n$. The outline for the sample-splitting procedure is as follows. Let the first n_1 observations form a training partition and the remaining $n_2 = n - n_1$ observations form a test partition. Using the training data, we generate a set of candidate detectors $\widehat{\mathcal{H}}_{+,\Lambda}$ across different hyperparameter settings Λ. Using the test data, the test statistic is defined as the maximum empirical score over the set of candidate detectors, i.e.

$$\hat{T}^{(split)}_{n,>} = \sup_{h\in\widehat{\mathcal{H}}_{+,\Lambda}} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (Y_i - \hat{p}_\delta(X_i))h(X_i). \tag{4}$$

We reject the null hypothesis if $\hat{T}^{(split)}_{n,>}$ exceeds critical value τ_α defined in the theorem below. Proofs for all the theoretical results are in the Appendix.

Theorem 1. *Let Y_i^* be the binary random variable with probability $\hat{p}_\delta(X_i)$. By setting τ_α to be the $1-\alpha$ quantile of*

$$T^{*(split)}_{>} = \sup_{h\in\widehat{\mathcal{H}}_{+,\Lambda}} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (Y_i^* - \hat{p}_\delta(X_i))h(X_i), \tag{5}$$

the Type I error of the sample-splitting test is controlled at level α.

Based on the theorem, we may use a simple Monte Carlo procedure to calculate τ_α which provides *assumption-free, finite-sample* Type I error control, in contrast to existing tests for model calibration. In particular, we construct bootstrap datasets $b = 1, \cdots, B$, where the outcomes $Y_i^{*(b)}$ are resampled with probability $\hat{p}_\delta(X_i)$, the probability distribution at the boundary of the null hypothesis space. Then the critical value is set to the $1-\alpha$ quantile of the the bootstrapped test statistics.

Given the Type I error guarantee, the next question is how to construct a set of detectors to maximize power (= 1−Type II error). As motivation, suppose we were only allowed to generate a single detector h. Per [Vaart, 1998], the local asymptotic power of the test with respect to h is determined by the ratio

$$\frac{\mathbb{E}\left[(Y - \hat{p}_\delta(X))h(X)\right]}{\sqrt{\mathrm{Var}\left((Y - \hat{p}_\delta(X))h(X)\right)}}. \tag{6}$$

Let $g_0(X) = p_0(X) - \hat{p}_\delta(X)$ denote the expected residuals. Given the constraint that detectors must be non-negative, the numerator is maximized by $g_0(X)\mathbb{1}\{g_0(X) \ge 0\}$. To maximize the ratio, we can tune over the broader class of detectors $h_{0,\gamma}(X) = g_0(X)\mathbb{1}\{g_0(X) > \gamma\}$ for $\gamma \ge 0$, which also reflects our interest in observations with the largest values of $g_0(X)$. This family of detectors also has a practical advantage. Because g_0 is unknown and we threshold on the estimated residuals in practice, the expected score will be large for some choice of γ as long as the estimated residual model is able to isolate some subset of observations with large expected residuals.

Given this motivation, we propose the following procedure for generating detectors. Suppose one has a set of candidate algorithms (e.g., random forests and neural networks) indexed by the set of hyperparameters Λ. For each $\lambda \in \Lambda$, we fit a residual model $\hat{g}_{\lambda,n}$ by training a regression model to predict the conditional mean of $\epsilon = Y - \hat{p}(X)$ given X using the training partition. We then construct the set of detectors

$$\hat{\mathcal{H}}_{+,\Lambda} = \left\{\hat{h}_{\lambda,\gamma,n} : \gamma \ge 0, \lambda \in \Lambda\right\} \tag{7}$$

where $\hat{h}_{\lambda,\gamma,n} = \hat{g}_{\lambda,n}(X)\mathbb{1}\{\hat{g}_{\lambda,n}(X) > \gamma\}$. The test statistic can now be rewritten as

$$\hat{T}^{(split)}_{n,>} = \max_{\lambda\in\Lambda} \underbrace{\max_{\gamma\ge 0} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (Y_i - \hat{p}_\delta(X_i))\, \hat{h}_{\lambda,\gamma,n}(X_i)}_{\text{Score-based CUSUM}}.$$

Notice that the inner summation corresponds exactly to the score-based cumulative sum (CUSUM) test statistic, which is typically used to detect changepoints along a single axis [Gombay, 2003, 2017, Feng et al.,

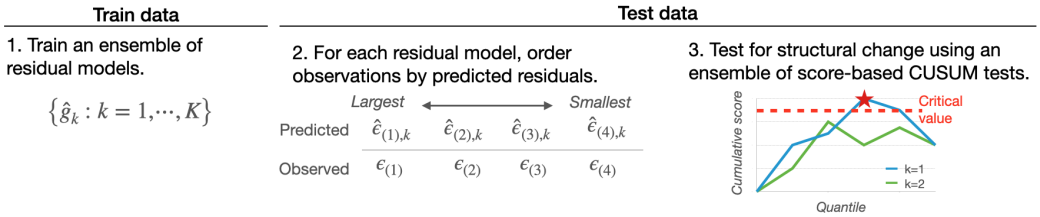


Figure 1: Summary of the procedure for the one-sided test. After training residual models, observations in the test partition are ordered by their predicted residuals. The poorly calibrated subgroup is detected using a changepoint test, as visualized in a control chart (right). The test statistic is the maximum cumulative score across all residual models (red star). The two-sided test orders observations by the absolute predicted residuals.

2022a]. So our procedure can be viewed as performing changepoint detection along data-adaptively defined axes $\{\hat{g}_{\lambda,n} : \lambda \in \Lambda\}$, where the null hypothesis is that the mean score is uniformly non-positive when observations are ordered by their predicted residuals from largest to smallest and the alternative is that the mean scores are positive prior to some changepoint γ and non-positive thereafter. Through this data-adaptive ordering, we expect the observed residuals with large positive values to aggregate at the beginning of this sequence as long as the residual model correctly identifies some region where the expected residuals $E[Y - \hat{f}(X)|X]$ indeed attain large positive values. Thus, the sequence of observed versus predicted residuals should initially be positively correlated, and the correlation should drop to zero outside of the correctly identified subset of poorly calibrated observations.

Leveraging this connection with the changepoint literature, we can visualize the test using "control charts," which are typically used to visualize changepoint detection procedures along a single dimension (Figure 1) [Montgomery, 2013, Feng et al., 2022b]. Under the alternative, when we plot the cumulative sum of the scores with respect to a residual model (X-axis can be interpreted as the index of the ordered observation or quantile of the predicted residual), we expect to observe a steady accumulation of positive residuals, resulting in a pronounced peak. The location of this peak indicates the size of the subgroup: an early peak means the identified subgroup is small whereas a very late peak means that the identified subgroup is large. Large positive slopes correspond to subgroups where the model is very poorly calibrated, whereas flat or negative slopes correspond to subgroups where model calibration is mostly within the desired tolerance. Thus the shape of the curve provides insight into the nature of the poorly calibrated subgroup. As the strong calibration test fits a suite of candidate residual models, we plot a curve for each residual model, and the maximum value attained across all curves corresponds to the test statistic.

Finally, one may ask (i) how many candidate residual models should one fit and (ii) how much data should one dedicate to training the residual models? These questions concern two tradeoffs. More residual models increase our chance of finding a changepoint but require more stringent multiplicity correction. Also, allocating more data to training can increase the accuracy of the residual models but reduces the sample size available for testing. We clarify the answer to these two questions in the following result. Note that c_k denote positive constants that depend on the variability of the residuals, the residual model classes, and their learning rates.

Theorem 2. *Suppose there is some* $\gamma \geq 0$ *such that*

$$\Psi_\gamma = \mathbb{E}\left[(Y - \hat{p}_\delta(X))\, h_{0,\gamma}(X)\right] > 0.$$

Consider any $\omega \leq 1$ *and* $\lambda \in \Lambda$. *For a sufficiently large* n_1 *such that*

$$\Psi_\gamma - c_1 n_1^{-\omega/2} \geq c_2 \sqrt{\frac{\log(|\Lambda|(n_2+1)/\alpha)}{n_2}},$$

statistical power $\Pr\left(\hat{T}^{(split)}_{n,>} > \tau_\alpha \mid \mathcal{S}^{(n_1)}_{\lambda,\omega}\right)$ *conditional on the event*

$$\mathcal{S}^{(n_1)}_{\lambda,\omega} = \left\{\mathbb{E}\left\|h_{0,\gamma}(X) - \hat{h}_{\lambda,\gamma,n}(X)\right\|^2 \leq c_3 n_1^{-\omega}\right\} \quad (8)$$

is lower bounded by

$$1 - \exp\left(-\frac{n_2\left(\Psi_\gamma - c_1 n_1^{-\omega/2} - c_2\sqrt{\log(|\Lambda|(n_2+1)/\alpha)/n_2}\right)^2}{2c_3}\right),$$

where $|\Lambda|$ *is the number of hyperparameters.*

Note that condition (8) is satisfied by any ML algorithm with a sufficiently fast convergence rate for appropriately chosen hyperparameters λ. Thus, the lower bound above states that the number of hyperparameters impacts power only through a logarithmic term, which justifies our approach of testing a suite of residual models. In contrast, existing methods test only a single residual model [Janková et al., 2020, Zhang et al., 2021], which is tuned using CV within the training partition. In settings with limited amounts of audit data, CV tends to overfit and select a suboptimal detector, leading to lower power. Moreover, this procedure is very computationally expensive

when combined with CV, because one would have to perform CV within CV.

Second, the lower bound grows much faster with the amount of test data than with the amount of training data. This highlights an interesting "phase change" in how much data one should allocate to training versus testing. One should allocate just enough training data so that (8) is satisfied with high probability (note that many ML algorithms have convergence rates of this form) and dedicate the rest to testing.

While Theorem 2 provides valuable intuition for how different hyperparameters of the procedure affect the power of the overall test, the constants in the bounds make it difficult to translate the results into practice. In the following section, we side-step the challenge of determining the optimal sample-splitting ratio by extending the procedure via cross-validation.

2.2 K-fold Cross-validation

We now extend the above procedure to use CV. This not only reduces the sensitivity of the procedure to the exact choice for the sample-splitting ratio but also improves power. The technical challenge is how to maintain Type I error control, despite the correlation between estimators across folds. Prior works provide only ad-hoc solutions [Zhang et al., 2021] or assume estimators converge to the oracle sufficiently fast Subbaswamy et al. [2021]. Here we present a procedure that requires very minimal assumptions.

We extend the sample-splitting procedure as follows. Partition the audit data into folds V_k for $k = 1, \ldots, K$. For each $\lambda \in \Lambda$, let $\hat{g}^{(-k)}_{\lambda,n}$ denote the estimated residual model using data in all but the k-th fold for $k = 1, \ldots, K$. The CV test statistic is defined as

$$\hat{T}^{(CV)}_{n,>} = \sup_{\lambda\in\Lambda,\gamma\geq 0} \frac{1}{|V_k|} \sum_{k=1}^{K} \sum_{(X_i,Y_i)\in V_k} (Y_i - \hat{p}_\delta(X_i)) \hat{h}^{(-k)}_{\lambda,\gamma,n}(X_i). \tag{9}$$

To establish Type I error control, we only require the weak assumption of uniform convergence.

Assumption 1. *For a given λ and γ, define $\bar{h}_{\lambda,\gamma,n}$ as the average detector estimated from $K-1$ folds of a dataset with n observations. That is, $\bar{h}_{\lambda,\gamma,n} = \mathbb{E}\left[\hat{g}^{(-1)}_{\lambda,n}(X)\mathbb{1}\left\{\hat{g}^{(-1)}_{\lambda,n}(X) \geq \gamma\right\}\right]$, where the expectation is with respect to the estimated residual models. Suppose that*

$$\sup_{\lambda\in\Lambda,\gamma\geq 0} \left\| \bar{h}_{\lambda,\gamma,n} - \hat{h}^{(-1)}_{\lambda,\gamma,n} \right\|_2 \to_p 0. \tag{10}$$

Under this assumption, we prove that one can essentially treat the estimated residual models as fixed and use the same Monte Carlo procedure as before to calculate the critical value. The only difference is that we control the Type I error rate asymptotically, rather than in finite samples.

Theorem 3. *Suppose Assumption 1 holds. Define Y^*_i using the definition in Theorem 1 and $T^{*(CV)}_{>}$ using (9) but replace Y_i with Y^*_i. If τ_α is set to the $1-\alpha$ quantile of $T^{*(CV)}_{>}$, the Type I error of the CV test is asymptotically controlled at level α.*

2.3 Extension to the two-sided test

Finally, we extend the procedure to test the two-sided null hypothesis. The key difference is that we now order observations by the magnitude of the predicted residuals, rather than their predicted residuals themselves. For ease of exposition, we only describe the sample-splitting procedure for the two-sided setting. The same ideas are used to extend the CV procedure.

Following the same logic as before, we begin with restating the two-sided hypothesis test in terms of a score test. Let $\mathcal{H}$ refer to the set of bounded functions, removing the prior restriction of non-negativity. For a given h, the working model for structural change is now $\text{logit}(p(X;h)) = \text{logit}(\hat{p}_{\delta\,\text{sign}(h)}) + \theta h(X)$, where $\text{sign}(h(X))$ is the sign of $h(X)$ and zero if $h(X) = 0$, and $\hat{p}_{\delta\,\text{sign}(h)}(X) = [\hat{p}(X) + \delta\,\text{sign}(h(X))]_{[0,1]}$. Thus we can reframe (1) as

$$\begin{aligned} H_0 &: \sup_{h\in\mathcal{H}} \mathbb{E}\left[\left(Y - \hat{p}_{\delta\,\text{sign}(h)}\right) h(X)\right] \leq 0 \\ H_1 &: \sup_{h\in\mathcal{H}} \mathbb{E}\left[\left(Y - \hat{p}_{\delta\,\text{sign}(h)}\right) h(X)\right] > 0. \end{aligned} \tag{11}$$

As before, we will instead perform a restricted score test by generating candidate detectors given a set of hyperparameters Λ. More specifically, for each $\lambda \in \Lambda$, we fit residual models $\hat{g}_{\lambda,n}(X)$ that estimate $p_0(X) - \hat{p}_\delta(X)$ if $p_0(X) > \hat{p}_\delta(X)$, $p_0(X) - \hat{p}_{-\delta}(X)$ if $p_0(X) < \hat{p}_{-\delta}(X)$, and zero otherwise. We then generate detectors $\hat{h}_{\lambda,\gamma,n}(x) = \hat{g}_{\lambda,n}(X)\mathbb{1}\{|\hat{g}_{\lambda,n}(X)| \geq \gamma\}$ for $\gamma \geq 0$. Consequently, the test statistic in the two-sided setting is the maximum of the score-based CUSUM statistics where observations are ordered by the absolute predicted residuals, i.e.

$$\hat{T}^{(split)}_n = \max_{\lambda\in\Lambda,\gamma\geq 0} \sum_{i=n_1+1}^{n} \left(Y_i - \hat{p}_{\delta\,\text{sign}(h_{\lambda,\gamma,n})(X_i)}\right) \hat{h}_{\lambda,\gamma,n}(X_i). \tag{12}$$

To calculate the critical value, we must modify the Monte Carlo procedure. Unlike the one-sided setting, it is no longer straightforward to determine the null distribution whose test statistic is stochastically largest, because estimated models may disagree on the sign of the expected residual. As such, we set the critical value to the quantile of a *modified* statistic that upper bounds (12). More specifically, for each X, we

sample *two* binary outcomes with marginal probabilities $\hat{p}_\delta(X)$ and $\hat{p}_{-\delta}(X)$. For each model $\hat{g}_{\lambda,n}$, we calculate a "bounding" CUSUM statistic by selecting the outcome generated with probability $\hat{p}_\delta(X)$ if the predicted residual is positive and $\hat{p}_{-\delta}(X)$ if the predicted residual is negative. The modified statistic is the maximum of these bounding CUSUM statistics. This procedure, formally described below, ensures finite-sample Type I error control.

Theorem 4. *Let U_i for $i = 1, \ldots, n$ be IID standard uniform random variables. Define $Y^*_{i,\lambda} = \mathbb{1}\{U_i \leq \hat{p}(X_i) + \delta \operatorname{sign}(\hat{g}_{\lambda,n}(X_i))\}$. Define $T^{*(split)}$ using (12) but replacing Y_i with $Y^*_{i,\lambda}$. Set the critical value τ_α to the $1-\alpha$ quantile of $T^{*(split)}$. For the two-sided hypothesis test, the sample-splitting procedure that rejects the null when $\hat{T}_n^{(split)} > \tau_\alpha$ controls the Type I error at level α.*

2.4 Variable importance plots

In addition to control charts, we can use variable importance (VI) plots to gain insight into potential reasons for model miscalibration. Here we consider a simple procedure using permutation VI to compute how important each variable is for detecting a poorly calibrated subgroup. (Future work may consider more sophisticated VI measures such as using Shapley values [Lundberg and Lee, 2017, Williamson and Feng, 2020].) For each variable, we permute its values and calculate the change in the test statistic. The importance of that variable is defined as the drop in the test statistic, where a larger drop indicates a more important variable. We emphasize that this definition of VI is *not* the same as ordinary VI measures that quantify how useful a variable is to a model's average performance. Ordinary VI measures describe the majority group and are not meant to characterize poorly calibrated subgroups.

3 SIMULATIONS

3.1 Setup

We begin with a simulation study comparing the proposed cross-validated procedure (`AdaptScoreCUSUM`) to existing tests for model calibration: the Hosmer-Lemeshow test (`ChiSq`) [Lemeshow et al., 2013], a score test based on Platt scaling (`Score`) [Platt, 1999], a score test based on the multicalibration procedure in [Kim et al., 2019] (`Multicalib`), an adaptive Chi-squared test that extends [Zhang et al., 2021] (`AdaptChiSq`), and an adaptive score test that extends [Janková et al., 2020] (`AdaptScoreSimple`). As shown in Table 1 of the Appendix, the procedures can be categorized based on whether they use prespecified versus adaptively-defined axes and whether they run implement Chi-squared versus score tests. All the tests that consider data-adaptive axes are assessing for strong calibration, while all the tests along prespecified axes assess for moderate calibration. We include the latter as they are commonly used in practice, even though they assess for a weaker form of calibration.

The three comparator score tests can be viewed as special cases of our procedure. `Score` implements a score-based CUSUM test with respect to the logistic recalibration model (2) along the prespecified axis $\text{logit}(\hat{p}(X))$. This test is not data-adaptive, so it does not require sample-splitting. `Multicalib` fits candidate residual models after performing a single split of the data and only considers detectors with thresholds fixed at $\gamma = 0$. `AdaptScoreSimple` further implements CV, so it is almost the same as ours except that the threshold is fixed at $\gamma = 0$. For the Chi-squared tests, we divided the data into 2 versus 10 bins. (Performance for other bin numbers was similar or worse.) Because Chi-squared tests are traditionally designed to test hypotheses with a tolerance of $\delta = 0$, we modified the test statistic to test non-zero tolerances.

For tests that prespecify axes, we follow standard practice and use the single axis $\hat{p}(X)$. For the data-adaptive tests, the procedures were unified under our framework to make them as comparable as possible. Residual models were fit using RFs and kernel logistic regression across various hyperparameter settings. The detectors used as input X and $\hat{p}(X)$. CV-based tests used 4 folds and single-split procedures reserved 25% of the data for testing. All methods used the proposed Monte Carlo procedure to calculate critical values and control Type I error.

Covariates $X \in \mathbb{R}^{10}$ were independently sampled from Uniform$[-5, 5]$. The outcome is sampled with the log odds as $(0.6x_0 + 0.4x_2 + 0.2x_3)\mathbb{1}\{\max(x_1, -x_2) \geq -2\} + 0.2x_1\mathbb{1}\{\max(x_1, -x_2) < -2\}$. The Appendix includes additional simulation details and results, as well as a simulation study verifying Type I error control.

We test the two-sided null hypothesis (1) with $\epsilon = 0$ for two algorithms: a logistic regression model (LR) that incorrectly assumes the logit is linear with respect to X and an RF. The RF is not misspecified but may converge slowly to the true risk. For tolerance levels $\delta = 0.025, 0.05$, and 0.075, the poorly calibrated subgroups had prevalences 0.6, 0.5, and 0.3 for LR and 0.8, 0.5, and 0.3 for RF, respectively.

3.2 Results

`AdaptScoreCUSUM` consistently outperformed other methods across all settings (Figure 2). Tests that prespecified the axis performed the worst. Adaptive

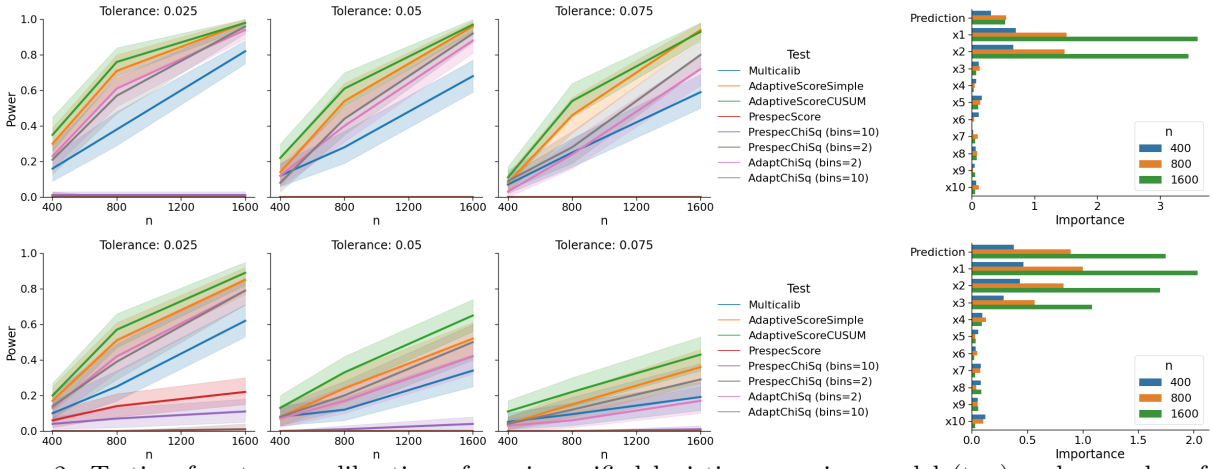


Figure 2: Testing for strong calibration of a misspecified logistic regression model (top) and a random forest model (bottom) across tolerance levels δ. Power is plotted against audit dataset sizes n on the left, where 95% confidence intervals are given by the shaded areas. Variable importance plots are on the right. Results are from 100 simulation replicates.

score-based tests attained much higher power than adaptive chi-squared tests, doubling it in certain settings. This improvement is particularly evident at higher tolerance levels, where the poorly calibrated subgroup is smaller and it becomes even more important to extract as much signal from the data as possible. `AdaptScoreCUSUM` offered the biggest improvements over `AdaptScoreSimple` in smaller sample sizes, where the residual models can be quite inaccurate. Thus the detectors used in `AdaptScoreSimple` with γ fixed to zero have difficulty isolating regions with poor calibration.

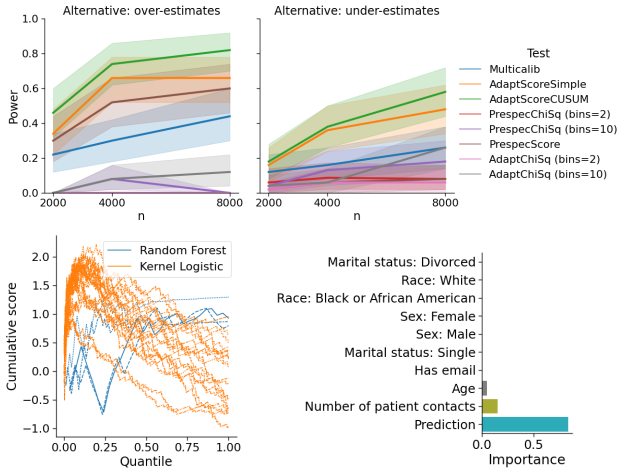
When auditing the LR model using `AdaptScoreCUSUM`, the test statistic was maximized by the RF-based detector in a majority of the cases. This is unsurprising, given that the subgroup structure in simulated data matches the structure learned by recursive partitioning. Moreover, the VI plots show that the detectors correctly recovered the misspecified subgroup, as variables $\hat{p}(x)$, and x_1, x_2 were assigned the highest importance. In contrast, when auditing the RF model, the `AdaptScoreCUSUM` test statistic was typically maximized by kernel LR. This illustrates how RFs are not very powerful for detecting miscalibration of an RF, because we have already extracted most of the signal from the data using recursive partitioning. Kernel LR uses an entirely different approach, so it is better at extracting the remaining signal. From the associated VI plots, we see that x_1, x_2, $\hat{p}(x)$, and x_3 are now the most important for detecting poor calibration. This likely reflects the fact that kernel LR converges faster than RF to the true probabilities in certain regions.

4 EMPIRICAL EXPERIMENTS

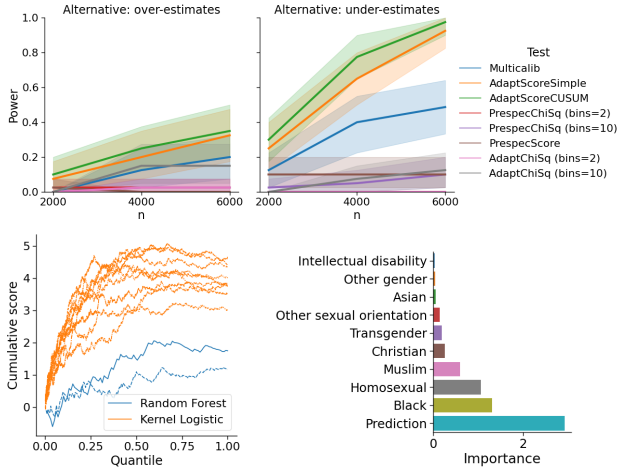
We now compare the procedures for auditing two binary classifiers trained on real-world data. The first model is an RF that predicts risk of 30-day unplanned readmission given data from the Electronic Health Records (EHR) from the Zuckerberg San Francisco General Hospital. We audit the model for strong calibration with respect to the demographic variables. The second model is a neural network (NN) that predicts whether social media comments are toxic, given data from the CivilComments dataset [Borkan et al., 2019] and embeddings extracted using a BERT model [Reimers and Gurevych, 2019]. We audited for strong calibration with respect to the demographic identities of each comment as well as the extracted embeddings. To differentiate between over- and under-estimation of the true risk, we tested the two one-sided null hypotheses separately. Additional details of the data analyses are in the Appendix.

Figure 3 shows that `AdaptScoreCUSUM` consistently achieved higher power than the other procedures. It detected over-estimation of the risk in the readmission model and under-estimation in the toxic comment classifier. There was only moderate evidence of miscalibration in the other direction for the two models. In the control chart for the readmission model, the cumulative scores increase for a short period and drop thereafter. This suggests that only a small subgroup is miscalibrated. In contrast, the control chart for the toxic comment classifier steadily trends upwards, suggesting that miscalibration is quite widespread.

VI plots provide further insight into the characteristics of the miscalibrated subgroups. For both models, the most important variable for identifying miscalibra-



(a) 30-day unplanned readmission model



(b) Toxic comment classifier

Figure 3: Auditing prediction models for strong calibration with tolerance $\delta = 0.05$. Top: Statistical power across audit dataset sizes n. Bottom left: Example control chart plotting cumulative score. Bottom right: importance of the top 10 variables

tion was the predicted risk from the model itself. The second most important variables for the readmission model and the toxic comment classifier were the number of patient contacts a patient had at the hospital (a good proxy for the social and medical needs of a patient) and whether the comment related to black identities, respectively. These results were further verified when we plotted calibration curves of the two models, stratified by the top variables (Figure 5 of the Appendix).

5 DISCUSSION

We have presented an adaptive score-based CUSUM procedure for testing if a given ML algorithm is poorly calibrated for some subgroup. The procedure is motivated by the idea that one can transform the problem of subgroup detection into the problem of changepoint detection by ordering observations by their predicted residuals. Along this sequence, we expect to see a change in the association between the observed and predicted residuals. This changepoint formulation lets us fully leverage the natural ordering of the data and the information learned by the estimated residual models, and avoid unnecessary binning of the data. As shown in the empirical experiments, the procedure consistently outperforms existing methods. The accompanying control charts and VI plots can also help users understand when a model is unreliable and inform model revision efforts.

Future work includes extending the current method to other types of outcomes (e.g. categorical and continuous) and non-tabular data. In addition, strong calibration is only one measure of fairness and only quantifies *statistical* bias, i.e. when the model fails to align with the data. This is separate from *societal* bias, which describes when the data fails to align with the desired state of the world. Different notions of fairness are useful in different contexts, so future investigations may consider how the proposed method can be extended to audit models for societal biases as well.

Acknowledgments

The authors are grateful to Adarsh Subbaswamy for helpful feedback and Lucas Zier, Jim Marks, and Steve Solnit for supplying the dataset from the Zuckerberg San Francisco General Hospital. This work was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award Center of Excellence in Regulatory Science and Innovation grant to University of California, San Francisco (UCSF) and Stanford University, U01FD005978. The contents are those of the author(s) and do not neces-

sarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

References

Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *J. Am. Med. Inform. Assoc.*, 28(3):549–558, March 2021. URL http://dx.doi.org/10.1093/jamia/ocaa283.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013. doi: 10.1287/mnsc.1120.1641. URL https://doi.org/10.1287/mnsc.1120.1641.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 491–500, New York, NY, USA, May 2019. Association for Computing Machinery.

R L Brown, J Durbin, and J M Evans. Techniques for testing the constancy of regression relationships over time. *J. R. Stat. Soc. Series B Stat. Methodol.*, 37 (2):149–192, 1975. URL http://www.jstor.org/stable/2984889.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, 17(7):392–406, July 2016. URL http://dx.doi.org/10.1038/nrg.2016.27.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1550–1553, April 2019. URL http://dx.doi.org/10.1109/ICDE.2019.00139.

D R Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565, 1958. URL http://www.jstor.org/stable/2333203.

Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 1467–1477, New York, NY, USA, August 2020. Association for Computing Machinery. URL https://doi.org/10.1145/3394486.3403199.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Oper. Res.*, September 2022. URL https://doi.org/10.1287/opre.2022.2363.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, June 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, January 2012. Association for Computing Machinery. URL https://doi.org/10.1145/2090236.2090255.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with Cross-Modal embeddings. *International Conference on Learning Representations*, May 2022. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Jean Feng, Alexej Gossmann, Gene Pennello, Nicholas Petrick, Berkman Sahiner, and Romain Pirracchio. Monitoring machine learning (ML)-based risk prediction algorithms in the presence of confounding medical interventions. November 2022a. URL http://arxiv.org/abs/2211.09781.

Jean Feng, Rachael V Phillips, Ivana Malenica, Andrew Bishara, Alan E Hubbard, Leo A Celi, and Romain Pirracchio. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *npj Digital Medicine*, 5(1):1–9, May 2022b. URL https://www.nature.com/articles/s41746-022-00611-y.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Inf Inference*, 10(2):455–482, June 2021. URL https://academic.oup.com/imaiai/article-pdf/10/2/455/38549621/iaaa017.pdf.

Pierre Glaser, David Widmann, Fredrik Lindsten, and Arthur Gretton. Fast and scalable score-based kernel calibration tests. *Conference on Uncertainty in Artificial Intelligence*, 216:691–700, 2023. URL https://proceedings.mlr.press/v216/glaser23a.html.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. January 2023. URL http://arxiv.org/abs/2301.13767.

David C. Goff, Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. O'Donnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Paul Sorlie, Neil J. Stone, and Peter W. F. Wilson. 2013 acc/aha guideline on the assessment of cardiovascular risk. *Circulation*, 129(25_suppl_2): S49–S73, 2014. doi: 10.1161/01.cir.0000437741.48606.98. URL https://www.ahajournals.org/doi/abs/10.1161/01.cir.0000437741.48606.98.

Edit Gombay. Sequential Change-Point detection and estimation. *Seq. Anal.*, 22(3):203–222, January 2003. URL https://doi.org/10.1081/SQA-120025028.

Edit Gombay. Editor's special invited paper: On the efficient score vector in sequential monitoring. *Sequential Analysis*, 36(4):435–466, October 2017. URL https://doi.org/10.1080/07474946.2017.1394728.

Parikshit Gopalan, Michael P Kim, and Omer Reingold. Characterizing notions of omniprediction via multicalibration. February 2023. URL http://arxiv.org/abs/2302.06726.

Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, February 2023. URL https://academic.oup.com/biomet/article-pdf/110/1/33/49160126/asac040.pdf.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

Douglas M Hawkins. Diagnostics for use with regression recursive residuals. *Technometrics*, 33(2):221–234, 1991. URL http://www.jstor.org/stable/1269048.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) masses. *International Conference on Machine Learning*, 80: 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

D W Hosmer, T Hosmer, S Le Cessie, and S Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Stat. Med.*, 16(9):965–980, May 1997. URL http://dx.doi.org/10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.0.co;2-o.

David W Hosmer and Nils Lid Hjort. Goodness-of-fit processes for logistic regression: simulation results. *Stat. Med.*, 21(18):2723–2738, September 2002. URL http://dx.doi.org/10.1002/sim.1200.

Aaron Hudson, Marco Carone, and Ali Shojaie. Inference on function-valued parameters using a restricted score test. May 2021. URL http://arxiv.org/abs/2105.06646.

Christina Ilvento. Metric learning for individual fairness. *Symposium on Foundations of Responsible Computing*, 2020. URL http://arxiv.org/abs/1906.00250.

Jana Janková, Rajen D Shah, Peter Bühlmann, and Richard J Samworth. Goodness-of-fit testing in high dimensional generalized linear models. *J. R. Stat. Soc. Series B Stat. Methodol.*, 82(3):773–795, July 2020. URL https://onlinelibrary.wiley.com/doi/10.1111/rssb.12371.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-Box Post-Processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 247–254, New York, NY, USA, January 2019. Association for Computing Machinery. URL https://doi.org/10.1145/3306618.3314287.

Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proc. Natl. Acad. Sci. U. S. A.*, 119(4), January 2022. URL http://dx.doi.org/10.1073/pnas.2108097119.

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-Cal: An optimal test for the calibration of predictive models. March 2022. URL http://arxiv.org/abs/2203.01850.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 76(1), 2014. URL https://rss-onlinelibrary-wiley-com.offcampus.lib.washington.edu/doi/10.1111/rssb.12021.

Stanley Lemeshow, Rodney X Sturdivant, and David W Hosmer, Jr. *Applied Logistic Re-*

gression. Wiley & Sons, Limited, John, 2013. URL https://openlibrary.org/books/OL38255075M/Applied_Logistic_Regression.

Mike Li, Hongseok Namkoong, and Shangzhou Xia. Evaluating model performance under worst-case subpopulations. *Conference on Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/908075ea2c025c335f4865f7db427062-Paper.pdf.

D Y Lin, L J Wei, and Z Ying. Model-checking techniques based on cumulative residuals. *Biometrics*, 58(1):1–12, March 2002. URL http://dx.doi.org/10.1111/j.0006-341x.2002.00001.x.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, pages 4765–4774, 2017.

Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: Metrics and recalibration. *Uncertain. Artif. Intell.*, 2022. URL https://openreview.net/pdf?id=BCg4lD8ice5.

Subha Maity, Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. Statistical inference for individual fairness. *International Conference on Learning Representations*, March 2021. URL http://arxiv.org/abs/2103.16714.

Charles Marx, Shengjia Zhao, Willie Neiswanger, and Stefano Ermon. Modular conformal calibration. June 2022. URL https://proceedings.mlr.press/v162/marx22a/marx22a.pdf.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Appl.*, 8(1):141–163, March 2021. URL https://doi.org/10.1146/annurev-statistics-042720-125902.

Douglas C Montgomery. *Statistical quality control*. John Wiley & Sons, Nashville, TN, 7 edition, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. URL https://www.researchgate.net/file.PostFileLoader.html?id=540479d7d11b8bb1588b459d&assetKey=AS%3A273601008209920%401442242971560.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, April 2020. URL https://hdsr.mitpress.mit.edu/pub/qedrwcz3/download/pdf.

Anian Ruoss, Mislav Balunović, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *Conference on Neural Information Processing Systems*, February 2020. URL https://proceedings.neurips.cc/paper/2020/file/55d491cf951b1b920900684d71419282-Paper.pdf.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR, 2021. URL https://proceedings.mlr.press/v130/subbaswamy21a.html.

Anastasios A Tsiatis. A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, 67(1):250–251, January 1980. URL https://academic.oup.com/biomet/article-pdf/67/1/250/6690321/67-1-250.pdf.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Ben Van Calster and Andrew J Vickers. Calibration of risk prediction models: impact on decision-analytic performance. *Med. Decis. Making*, 35(2): 162–169, February 2015. URL http://dx.doi.org/10.1177/0272989X14547233.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.*, 74:167–176, June 2016. URL http://dx.doi.org/10.1016/j.jclinepi.2015.12.005.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92(2):349–376, September 2013. URL https://doi.org/10.1007/s10994-013-5355-6.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Boston, MA, 2005. URL https://link.springer.com/book/10.1007/b106715.

Vladimir Vovk, Ivan Petej, Paolo Toccaceli, Alexander Gammerman, Ernst Ahlberg, and Lars Carlsson. Conformal calibrators. *Symposium on Conformal and Probabilistic Prediction and Applications*, 128:84–99, 2020. URL https://proceedings.mlr.press/v128/vovk20a.html.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, February 2019. URL https://play.google.com/store/books/details?id=IluHDwAAQBAJ.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Conference on Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/file/1c336b8080f82bcc2cd2499b4c57261d-Paper.pdf.

Brian D Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. *International Conference on Machine Learning*, 2020. URL https://proceedings.icml.cc/static/paper_files/icml/2020/3042-Paper.pdf.

Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. Auditing ML models for individual bias and unfairness. *International Conference on Artificial Intelligence and Statistics*, March 2020. URL http://proceedings.mlr.press/v108/xue20a/xue20a.pdf.

Jiawei Zhang, Jie Ding, and Yuhong Yang. Is a classification procedure good Enough?—A Goodness-of-Fit assessment tool for classification learning. *J. Am. Stat. Assoc.*, pages 1–11, September 2021. URL https://doi.org/10.1080/01621459.2021.1979010.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In Hal Daumé Iii and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11387–11397. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhao20e.html.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No. This method can be executed using the CPU of an ordinary laptop]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A EXTENSION: TESTING NONZERO ϵ

To test for poorly calibrated subgroups with some minimum prevalence $\epsilon > 0$, the only modification needed is to constrain the set of detectors to those for which

$$\mathbb{E}\left[\mathbb{1}\{h(X) > 0\}\right] > \epsilon. \tag{13}$$

So we can use essentially the same sample-splitting or CV procedure, except we only consider thresholds γ for which the corresponding detector satisfies (13).

B PROOFS

Below we present proofs for all the theoretical results in the main manuscript. We use c_k to denote positive constants.

Proof of Theorem 1

Proof. It suffices to prove that conditional on the training data, the test statistic for the distribution with conditional probabilities equal to $\hat{p}_\delta$ stochastically dominates the test statistic for any other distribution under the null with conditional probabilities equal to p_0. To do this, we use a coupling argument.

Given any x, we can generate binary random variables (RVs) Y and $\tilde{Y}$ where $\Pr(Y = 1|x) = p_0(x)$, $\Pr(\tilde{Y} = 1|x) = \hat{p}_\delta(x)$, and $Y \leq \tilde{Y}$ as follows. First, sample a standard uniform random variable U. Then let $Y = \mathbb{1}\{U \leq p_0(x)\}$ and $\tilde{Y} = \mathbb{1}\{U \leq \hat{p}_\delta(x)\}$. As such, the above conditions are satisfied.

Using this procedure, we can generate coupled outcomes for observations in the test partition (i.e. $i = n_1 + 1, \cdots, n$). Consequently, the test statistics on the coupled test data must satisfy

$$\hat{T}_n^{(split)} = \sup_{h \in \widehat{\mathcal{H}}_{+,\Lambda}} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (Y_i - \hat{p}_\delta(X_i))h(X_i) \leq \tilde{T}_n^{(split)} = \sup_{h \in \widehat{\mathcal{H}}_{+,\Lambda}} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (\tilde{Y}_i - \hat{p}_\delta(X_i))h(X_i).$$

As such, $\hat{T}_n^{(split)}$ stochastically dominates $\tilde{T}_n^{(split)}$. □

Proof for Theorem 2

Proof. Below, we use the $\mathbb{P}_{n_1+1:n}$ and $\mathbb{P}$ to denote the empirical average over the test data split and the expectation, respectively.

We begin with determining the minimum value of τ_α to control Type I error. In particular, we must perform a multiplicity correction to account for the multiple residual models being tested. By a union bound, we have that

$$\Pr\left(\max_{h \in \widehat{H}_{+,\Lambda}} \left(\mathbb{P}_{n_1+1:n} - \mathbb{P}\right)(Y - \hat{p}_\delta(X))h(X) > \tau_\alpha\right) \tag{14}$$

$$\leq |\Lambda| \max_{\lambda \in \Lambda} \Pr\left(\sup_{\gamma \geq 0} \left(\mathbb{P}_{n_1+1:n} - \mathbb{P}\right)(Y - \hat{p}_\delta(X))\hat{g}_{\lambda,n}(X)\mathbb{1}\left\{\hat{g}_{\lambda,n} \geq \gamma\right\} > \tau_\alpha\right). \tag{15}$$

Applying Theorem 4.10 in [Wainwright, 2019], we have for any $\lambda \in \Lambda$ and $b \geq 0$ that

$$\Pr\left(\sup_{\gamma \geq 0} \left(\mathbb{P}_{n_1+1:n} - \mathbb{P}\right)(Y - \hat{p}_\delta(X))\hat{g}_{\lambda,n}(X)\mathbb{1}\left\{\hat{g}_{\lambda,n} \geq \gamma\right\} > 2\mathcal{R}_\lambda + b\right) \leq \exp\left(-\frac{n_2 b^2}{2c_1^2}\right) \tag{16}$$

where $\mathcal{R}_\lambda$ is an upper-bound of the Rademacher complexity for the function class

$$\left\{x \mapsto \hat{h}_{\lambda,\gamma,n}(x) = \hat{g}_{\lambda,n}(x)\mathbb{1}\left\{\hat{g}_{\lambda,n}(x) \geq \gamma\right\} : \gamma \geq 0\right\}.$$

Because the set of functions $\{x \mapsto \mathbb{1}\{\hat{g}_{\lambda,n}(x) > \gamma\} : \gamma \geq 0\}$ has VC dimension 1, we have by an application of Lemma 4.14 in [Wainwright, 2019] that

$$\mathcal{R}_\lambda \leq c_2 \sqrt{\frac{\log(n_2+1)}{n_2}}. \tag{17}$$

Plugging (17) and (16) into (15), we find that by setting

$$\tau_\alpha \geq c_3 \sqrt{\frac{\log(|\Lambda|(n_2+1)/\alpha)}{n_2}}, \tag{18}$$

controls the finite-sample Type I error at level α.

Now suppose n_1 is chosen so that $\Psi_\gamma - c_3 n_1^{-\omega/2} \geq \tau_\alpha$. Note that by Cauchy Schwarz, we have that

$$\left| \mathbb{P}\left[\left((Y - \hat{p}_\delta(X)) \left(\hat{h}_{\lambda,\gamma,n}(X) - h_{0,\gamma}(X) \right) \right] \right| \leq \; c_4 \sqrt{\mathbb{P}\left\| \hat{h}_{\lambda,\gamma,n}(X) - h_{0,\gamma}(X) \right\|^2}.$$

So conditional on the set $\mathcal{S}^{(n_1)}_{\lambda,\omega}$, the difference in the expected score is no greater than $c_5 n_1^{-\omega/2}$ for some $c_5 > 0$. Because $((Y - \hat{p}_\delta(X))h_{0,\gamma}(X)$ is sub-gaussian, we have by Chernoff's bound that

$$\begin{aligned}
&\Pr\left(\hat{T}_n^{(split)} < \tau_\alpha | \mathcal{S}^{(n_1)}_{\lambda,\omega} \right) \\
\leq &\Pr\left(\left(\mathbb{P}_{n_1+1:n} - \mathbb{P}\right) \left(Y - \hat{p}_\delta(X)\right) \hat{h}_{\lambda,\gamma,n}(X) + \mathbb{P}\left(Y - \hat{p}_\delta(X)\right) \left(\hat{h}_{\lambda,\gamma,n}(X) - h_{0,\gamma}(X) \right) < \tau_\alpha - \Psi_\gamma \right) \\
\leq &\exp\left(- \frac{n_2 \left(\Psi_\gamma - c_5 n_1^{-\omega/2} - \tau_\alpha \right)^2}{c_6} \right).
\end{aligned}$$

Plugging in (18) to the above expression gives us our desired result. □

Proof of Theorem 3

Proof. We will use $\mathbb{P}_{n,k}$ to denote the empirical mean with respect to fold k for $k = 1, \cdots, K$ and $\mathbb{P}_{n,-k}$ denote the mean with respect to data from all but the kth fold. Consider the decomposition

$$\sqrt{n} \begin{pmatrix} \mathbb{P}_{n,1}\left(Y - \hat{p}_\delta(X)\right) \hat{h}^{(-1)}_{\lambda,\gamma,n}(X) \\ \vdots \\ \mathbb{P}_{n,K}\left(Y - \hat{p}_\delta(X)\right) \hat{h}^{(-K)}_{\lambda,\gamma,n}(X) \end{pmatrix} = \sqrt{n} \begin{pmatrix} \mathbb{P}_{n,1}\left(Y - \hat{p}_\delta(X)\right) \bar{h}_{\lambda,\gamma,n}(X) \\ \vdots \\ \mathbb{P}_{n,K}\left(Y - \hat{p}_\delta(X)\right) \bar{h}_{\lambda,\gamma,n}(X) \end{pmatrix} \tag{19}$$

$$+ \sqrt{n} \begin{pmatrix} \left(\mathbb{P}_{n,1} - \mathbb{P}\right)\left(Y - \hat{p}_\delta(X)\right) \left(\hat{h}^{(-1)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \right) \\ \vdots \\ \left(\mathbb{P}_{n,K} - \mathbb{P}\right)\left(Y - \hat{p}_\delta(X)\right) \left(\hat{h}^{(-K)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \right) \end{pmatrix} \tag{20}$$

$$+ \sqrt{n} \begin{pmatrix} \mathbb{P}\left(Y - \hat{p}_\delta(X)\right) \left(\hat{h}^{(-1)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \right) \\ \vdots \\ \mathbb{P}\left(Y - \hat{p}_\delta(X)\right) \left(\hat{h}^{(-K)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \right) \end{pmatrix}. \tag{21}$$

First, we show that (21) is equal to zero. To see this, we have by the law of iterated expectations that

$$\mathbb{P}\left(Y - \hat{p}_\delta(X)\right) \left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \right) = \mathbb{P}_{n,-k} \left[\mathbb{P}\left[Y - \hat{p}_\delta(X) \mid X \right] \left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \right) \right], \tag{22}$$

where we marginalize over data in the kth fold. By the definition of $\bar{h}_{\lambda,\gamma,n}$, we have that $\mathbb{P}[\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X) \mid X = x] = 0$ for all x if we marginalize over all but the kth fold data (i.e. the data used to learn $\hat{h}^{(-k)}_{\lambda,\gamma,n}$). So the right hand side of (22) is equal to zero.

Next, we show that (20) is $o_p(1)$. Because the class of detectors varies across n, we will apply Theorem 19.28 in [Vaart, 1998], which is a generalization of Donsker's theorem that allows the indexing class to vary over n. To apply this result, note that the Lindeberg condition is satisfied, as residuals and detectors in the set $\widehat{H}_\Lambda$ are bounded. In addition, the bracketing entropy requirements are also satisfied. Thus we have that the stochastic process

$$\left\{(\lambda,\gamma) \mapsto \sqrt{n}\left(\mathbb{P}_{n,k} - \mathbb{P}\right)\left(Y - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X)\right)\right\} \tag{23}$$

converges to a mean-zero Gaussian process with covariance function $\Sigma((\lambda_1,\gamma_1),(\lambda_2,\gamma_2))$ equal to

$$\lim_n \mathrm{Cov}\left(\left(Y - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda_1,\gamma_1,n}(X) - \bar{h}_{\lambda_1,\gamma_1,n}(X)\right), \left(Y - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda_2,\gamma_2,n}(X) - \bar{h}_{\lambda_2,\gamma_2,n}(X)\right)\right). \tag{24}$$

By Assumption 1, (24) converges to zero. Thus we have that

$$\sqrt{n}\begin{pmatrix}\left(\mathbb{P}_{n,1} - \mathbb{P}\right)\left(Y - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X)\right)\\ \vdots \\ \left(\mathbb{P}_{n,K} - \mathbb{P}\right)\left(Y - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X)\right)\end{pmatrix} = o_p(1) \tag{25}$$

Combining the above results, we have established that

$$\sup_{\lambda,\gamma}\sqrt{n}\left(\sum_{k=1}^K \mathbb{P}_{n,k}\left(Y - \hat{p}_\delta(X)\right)\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \sum_{k=1}^K \mathbb{P}_{n,k}\left(Y - \hat{p}_\delta(X)\right)\bar{h}_{\lambda,\gamma,n}(X)\right) = o_p(1). \tag{26}$$

Having established that the remainder terms (20) and (21) are negligible, we can use the same arguments used to prove Theorem 1 to prove that the setting the critical value τ_α to the $1-\alpha$ quantile of

$$T^{*(CV,oracle)}_{n,>} := \sup_{\lambda\in\Lambda,\gamma\geq 0}\frac{1}{|V_k|}\sum_{k=1}^K \sum_{(X_i,Y_i^*)\in V_k}\left(Y_i^* - \hat{f}_\delta(X_i)\right)\bar{h}_{\lambda,\gamma,n}(X_i), \tag{27}$$

where Y_i^* is a resampled binary RV that is equal to one with probability $\hat{p}_\delta(X_i)$, controls the Type I error asymptotically.

In practice, $\bar{h}_{\lambda,\gamma,n}$ is unknown. So instead, we calculate the quantile for $T^{*(CV)}_{n,>}$, which plugs in the estimated detectors instead. To prove that this plug-in approach maintains asymptotic Type I error control, we must show that the difference between $T^{*(CV)}_{n,>}$ and $T^{*(CV,oracle)}_{n,>}$ is asymptotically negligible. Let $\mathbb{P}^*_{n,k}$ denote the empirical mean in the k-th fold with respect to the resampled outcomes $(Y_1^*,\cdots,Y_n^*)$. Similarly, let $\mathbb{P}^*$ denote the expectation with respect to the distribution with conditional probability equal to $\hat{p}_\delta$. Consider the following decomposition for each $k = 1,\cdots,K$:

$$\sqrt{n}\mathbb{P}^*_{n,k}\left(Y^* - \hat{f}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X)\right) = \sqrt{n}\left(\mathbb{P}^*_{n,k} - \mathbb{P}^*\right)\left(Y^* - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X)\right) \tag{28}$$

$$+\ \sqrt{n}\mathbb{P}^*\left(Y - \hat{p}_\delta(X)\right)\left(\hat{h}^{(-k)}_{\lambda,\gamma,n}(X) - \bar{h}_{\lambda,\gamma,n}(X)\right). \tag{29}$$

Using the same arguments as above, we have that (28) is $o_p(1)$ by Theorem 19.28 in [Vaart, 1998] and (29) is equal to zero. Summing these results over all k, we have established that the scaled difference between the calculated and oracle test statistic $\frac{1}{\sqrt{n}}\left(T^{*(CV)}_{n,>} - T^{*(CV,oracle)}_{n,>}\right)$ is $o_p(1)$. Therefore, by Slutsky's theorem, the $1-\alpha$ quantile for $T^{*,(CV)}_n$ controls the Type I error at the desired rate. □

Proof of Theorem 4

Proof. We again use a coupling argument to prove that the modified statistic stochastically dominates the test statistic under any distribution p_0 satisfying the null hypothesis. In particular, consider the sampling procedure where U is a standard uniform RV, $Y = \mathbb{1}\{U \leq p_0(X)\}$, $Y^{(-1)} = \mathbb{1}\{U \leq \hat{p}_{-\delta}(X)\}$, and $Y^{(1)} = \mathbb{1}\{U \leq \hat{p}_\delta(X)\}$.

	Chi-squared tests	Score-based tests
Prespecified axis	• Hosmer-Lemeshow: `ChiSq`	• Score test for Platt scaling: `Score`
Data-adaptive axes	• Based on [Zhang et al., 2021]: `AdaptChiSq`	• Based on [Kim et al., 2019]: `Multicalib` • Based on [Janková et al., 2020]: `AdaptScoreSimple` • Proposed: `AdaptScoreCUSUM`

Table 1: Categorization of existing and proposed testing procedures

Thus the conditional probabilities $\Pr(Y = 1|X)$, $\Pr(Y^{(-1)} = 1|X)$, and $\Pr(Y^{(1)} = 1|X)$ are $p_0(X)$, $\hat{p}_{-\delta}(X)$, and $\hat{p}_{\delta}(X)$, respectively. Also, for any h and i, we have that

$$\left(Y - \hat{p}_{\delta\,\mathrm{sign}(\hat{h}_{\lambda,\gamma,n})}(X)\right)\hat{h}_{\lambda,\gamma,n}(X) \leq \left(Y^{(\mathrm{sign}(\hat{h}_{\lambda,\gamma,n}))} - \hat{p}_{\delta\,\mathrm{sign}(\hat{h}_{\lambda,\gamma,n})}(X)\right)\hat{h}_{\lambda,\gamma,n}(X). \tag{30}$$

So if we use this procedure to resample the binary outcomes for the test partition, we would have that

$$\max_{\lambda\in\Lambda,\gamma\geq 0}\sum_{i=n_1+1}^{n}\left(Y_i - \hat{p}_{\delta\,\mathrm{sign}(\hat{h}_{\lambda,\gamma,n})}(X_i)\right)\hat{h}_{\lambda,\gamma,n}(X_i) \leq \max_{\lambda\in\Lambda,\gamma\geq 0}\sum_{i=n_1+1}^{n}\left(Y^{(\mathrm{sign}(\hat{h}_{\lambda,\gamma,n}))} - \hat{p}_{\delta\,\mathrm{sign}(\hat{h}_{\lambda,\gamma,n})}(X_i)\right)\hat{h}_{\lambda,\gamma,n}(X_i). \tag{31}$$

This implies our desired result. □

C ADDITIONAL SIMULATION DETAILS

For the residual models, we fit random forests and kernel logistic regression using the scikit-learn package [Pedregosa et al., 2011]. We tuned the following hyperparameters for RF: maximum number of features $p = 5$ versus $p = 10$ and max depth of 4 versus 8. For kernel logistic regression, we used an approximation of the polynomial kernel with degree two and subsequently fit a ridge-penalized logistic regression model, where we considered a regularization factor of $C = 1000, 100$, and 10.

D SIMULATION STUDY OF TYPE I ERROR CONTROL

The goal of this simulation is to analyze the Type I error of the score-test in finite samples. We test for strong calibration with tolerance $\delta = 0.025$. For the original ML algorithm, we train a logistic regression model using 10,000 observations generated with the conditional log odds as

$$\mathrm{logit}(p_{\mathrm{orig}}(X)) = 0.6x_1 + 0.4x_2 + 0.2x_3.$$

For the audit data, we simulated outcomes where the conditional probabilities were $p_{\mathrm{orig}} + 0.025$ to maximize Type I error. This simulation also reflects situations where ML algorithm is developed for one context and deployed in another, and one is interested in knowing if the ML algorithm is miscalibrated in some subgroup in the new target population.

We perform a one-sided test to determine if the predicted risks are under-estimates and a two-sided test. We implement the CV-based testing procedures for which we have only established asymptotic control of the Type I error rate. We do not include results from the single sample split, since we proved that it provides finite sample control of the Type I error rate. The critical values were set to target a Type I error rate of 0.1. Recall that the one-sided test calculates the critical value by sampling from the single worst-case null distribution, whereas the two-sided test relies on sampling upper bounds of the test statistic. As such, we expect the observed Type I error rate to be lower (i.e. more conservative) for the two-sided test than the one-sided test. A total of 100 simulation replicates were run.

As shown in Figure 4, we observe that Type I error is controlled across a variety of audit dataset sizes, including $n = 100$. So, even though our procedure only guarantees finite sample error rate control for the sample-splitting version, we are able to maintain Type I error control for small sample sizes even in the CV version because the assumptions needed for the CV procedure are very weak.

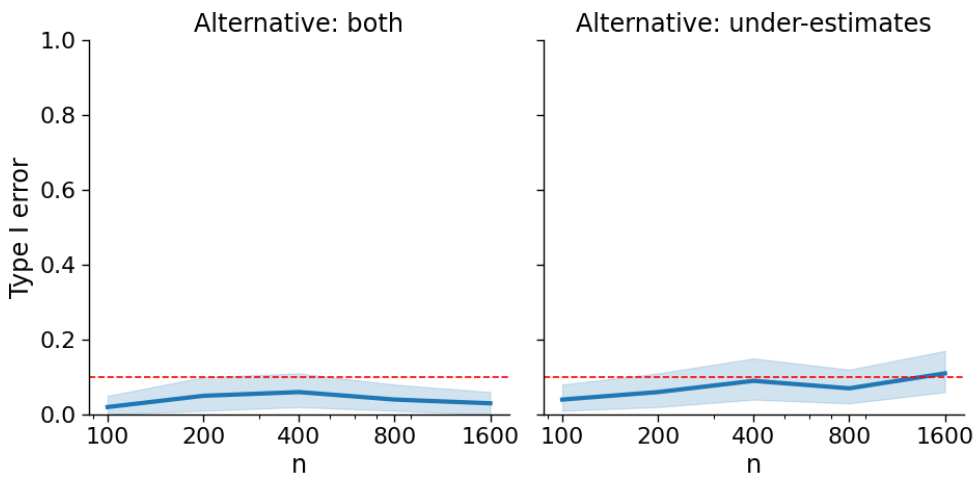
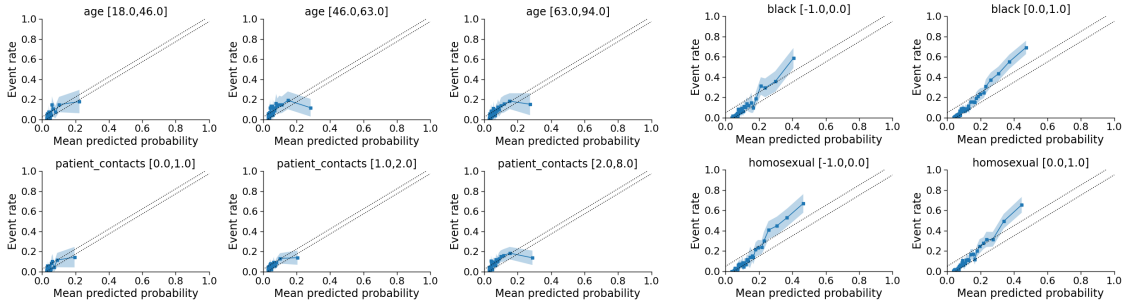


Figure 4: Simulation study of Type I error control. Note that the x–axis is shown on the log scale. Shaded areas correspond to 95% confidence intervals.



(a) Readmission model

(b) Toxic comment model

Figure 5: Calibration plots, stratified by most important demographic variables

E ADDITIONAL DATA ANALYSIS DETAILS

ZSFG dataset details: This is a private dataset provided by ZSFG under IRB approval. The entire dataset is composed of 88400 patients. For each replicate, we randomly selected 22,000 patients to train an RF to predict 30-day unplanned readmission risk and randomly selected $n = 2000, 4000, 8000$ observations for testing strong calibration. Use of this dataset was approved by the ZSFG Institutional Review Board.

Civil Comments dataset details: This public dataset [Borkan et al., 2019] is composed of 440,000 comments. For each replicate, we randomly selected 8000 comments to train a dense neural network and randomly selected $n = 2000, 4000, 6000$ observations to audit for strong calibration.