A structured regression approach for evaluating model performance across intersectional subgroups

Christine Herlihy*
cherlihy@umd.edu
University of Maryland
College Park, USA

Kimberly Truong*
truonkim@oregonstate.edu
Oregon State University
Corvallis, USA

Alexandra Chouldechova
alexandrac@microsoft.com
Microsoft Research
New York, USA

Miroslav Dudík
mdudik@microsoft.com
Microsoft Research
New York, USA

ABSTRACT

Disaggregated evaluation is a central task in AI fairness assessment, where the goal is to measure an AI system's performance across different subgroups defined by combinations of demographic or other sensitive attributes. The standard approach is to stratify the evaluation data across subgroups and compute performance metrics separately for each group. However, even for moderately-sized evaluation datasets, sample sizes quickly get small once considering intersectional subgroups, which greatly limits the extent to which intersectional groups are included in analysis. In this work, we introduce a structured regression approach to disaggregated evaluation that we demonstrate can yield reliable system performance estimates even for very small subgroups. We provide corresponding inference strategies for constructing confidence intervals and explore how goodness-of-fit testing can yield insight into the structure of fairness-related harms experienced by intersectional groups. We evaluate our approach on two publicly available datasets, and several variants of semi-synthetic data. The results show that our method is considerably more accurate than the standard approach, especially for small subgroups, and demonstrate how goodness-of-fit testing helps identify the key factors that drive differences in performance.

CCS CONCEPTS

• **Mathematics of computing** → Regression analysis; • **General and reference** → Evaluation; Measurement.

*Both authors contributed equally to this research.

1 INTRODUCTION

A core task when assessing the fairness of an AI system is measuring its performance across different subgroups defined by combinations of demographic or other sensitive attributes. Many of the best-known studies of algorithmic bias are grounded in this type of analysis [1, 5, 21, 25, 30], including Buolamwini and Gebru's Gender Shades study [5], which found that commercial gender classifiers have much higher error rates for darker-skinned women than other groups, and Obermeyer et al.'s study [25] finding bias in commercial algorithms used to guide healthcare decisions.

In their work formalizing this type of analysis, Barocas et al. [3] introduce the term *disaggregated evaluation* to refer to this task. The authors draw attention to the many decisions that implicitly shape any given disaggregated evaluation: from who will be involved, to what data will be used, which statistical approach taken, and what kinds of inferences drawn. In our work, we focus on the question of statistical methodology given an available dataset and pre-determined subgroups and performance metrics of interest. We introduce a method for estimating performance across subgroups that we show (i) is more accurate than approaches taken in standard practice; and (ii) can provide greater insight into which factors drive observed variation in performance. We do so through careful adaptations of well-established techniques rather than development of entirely novel statistical methods.

The "standard approach" to disaggregated evaluation proceeds by stratifying the evaluation data across subgroups and then conducting inference (i.e., computing performance metrics, confidence intervals, or other statistics) separately for each group. The primary challenge when applying this approach comes from small sample sizes. Even for moderately-sized evaluation datasets, sample sizes quickly get small once considering intersectional subgroups. For instance, in a medical diabetes mellitus dataset we use later in the paper, we have a 5000-patient evaluation dataset, of which 2689 patients are female, 620 are female *and* over age 80, but only 6 are female, over age 80, *and* Hispanic. Indeed, of the 32 distinct gender-age-race/ethnicity subgroups that can be formed in the data, 8 (i.e., 25%) have fewer than 10 observations, and nearly half have fewer than 25 observations. Inference based on so few observations is often uninformative, and may be unreliable. In practice, subgroups that are too small tend to be either excluded from analysis or merged with other small but potentially heterogeneous subgroups to form higher-level "catch-all" categories (e.g., "other"). These practices

Christine Herlihy, Kimberly Truong, Alexandra Chouldechova, and Miroslav Dudík

greatly limit the extent to which intersectional groups are even considered in many disaggregated evaluations. As a consequence, standard assessments may fail to surface fairness-related harms that could disproportionately affect intersectional subgroups [7], which in turn means that steps to mitigate those harms might not be taken.

In this work, we introduce a structured regression approach to disaggregated evaluation that we demonstrate can yield reliable system performance estimates even for very small subgroups (e.g., for groups with fewer than 25 observations). We also provide corresponding inference strategies for constructing confidence intervals for the subgroup-level performance estimates. We then demonstrate how goodness-of-fit testing can provide insight into the structure of fairness-related harms experienced by intersectional groups and also identify situations where observed variation in performance is attributable to benign factors. Lastly, we present results on two publicly available datasets, and several variants of semi-synthetic data. The results show that our method is considerably more accurate than the standard approach, especially for small subgroups. They further show that our method outperforms more statistically sophisticated baselines, including the *model-based metrics* method introduced by Miller et al. [23], while also offering additional advantages. We conclude by discussing limitations and future directions.

2 BACKGROUND AND RELATED WORK

In their taxonomy of sociotechnical harms of algorithmic systems, Shelby et al. [27] identify five high-level categories of harm: representational, allocative, quality of service, interpersonal, and social system. Our work contributes to the broader literature characterizing and assessing *allocative* and *quality-of-service* harms that can result from the use of algorithmic systems. Allocative harms, first discussed by Barocas et al. [2], occur when systems produce an inequitable distribution of information, opportunity, or resources across groups. As a running example, we consider a hypothetical setting in which a model trained to predict 30-day hospital readmission is used to prioritize high-risk patients for more intensive post-discharge care. Allocative harms might occur if certain subgroups of patients are disproportionately under-prioritized for more intensive care (i.e., have low selection rates) or are under-selected relative to their observed rate of readmission (i.e., have high false negative rates).

Quality-of-service harms occur when algorithmic systems underperform for certain socially salient groups of users [27, 35]. We examine quality-of-service harms across race and gender groups in the context of commercial automated speech recognition (ASR) systems using data previously analyzed by Koenecke et al. [21]. Specifically, we assess whether there is significant variation in the word error rate (WER) of the ASR systems across intersectional race and gender subgroups.

The term "intersectionality" was introduced by Crenshaw [7] to describe the distinct patterns of discrimination and disadvantage experienced by Black women, which she argued cannot be understood in terms of race or gender discrimination alone. In recent years, algorithmic fairness research has examined intersectional bias from many perspectives. This includes work introducing quantitative metrics intended to capture notions of intersectional fairness, such

as subgroup fairness [20], differential fairness [11, 12], and multi-calibration [14], along with learning algorithms for estimating and achieving these criteria. Wang et al. [34] study "predictivity differences" across intersectional subgroups, and discuss limitations of existing summary statistics (such as the maximum disparity across all groups) in capturing meaningful notions of intersectional harm. Our work differs from this literature because we are specifically interested in the task of disaggregated evaluation. This entails estimating and reporting system performance for each intersectional subgroup, rather than computing a particular fairness metric or learning a fairness-constrained model.

Our work most directly contributes to the growing literature introducing more sample-efficient methods for conducting disaggregated evaluations. This literature includes methods that leverage unlabelled data in model evaluation [6, 18, 19]; methods that bound or approximate performance for intersectional subgroups using marginal statistics [24]; and synthetic data augmentation approaches [32]. In work more closely related to the spirit of our structured regression approach, Piratla et al. [26] introduce the *attributed accuracy assay* (AAA) method, which models the accuracy of a model as a function of sensitive attributes and other features via a Gaussian process (GP). While we do not rely on GPs, we do proceed similarly by modeling the accuracy (or error) of a given model. Whereas we are specifically concerned with fairness and disaggregated evaluation, Piratla et al. [26] aim to produce an "accuracy surface" model that clients can use to estimate the performance of an existing model on their data.

The most closely related work in recent literature is that of Miller et al. [23], who introduce a Bayesian structured regression approach that they call *model-based metrics* (MBM). Their method applies to AI models that produce a score (say to predict a risk of hospital readmission). By modeling the distribution of scores given select features and the observed outcome, they are able to make inference on any performance metric of interest, but the approach is not directly applicable to the evaluation of models that do not produce classification scores (e.g., MBM does not directly apply to the evaluation of WER in ASR systems). Unlike the MBM approach, we model the target metric directly and fit separate models for each performance metric of interest. Our experiments show that our method yields more accurate estimates than MBM (see §5.1).

Our approach is also related to the classical line of research on normal means estimation, originating with the James-Stein (JS) estimator [15, 28]. The JS estimator works by shrinking standard estimates towards zero (or some other constant). This leads to a substantial decrease in variance, which outweighs a moderate increase in bias, and yields a more accurate estimator. The empirical Bayes (EB) approach [8] also leads to a form of shrinkage, but its motivation is different. It posits a hierarchical Bayesian model and estimates metric values by posterior means, while fixing prior hyperparameters to their point estimates. Our estimator works by optimizing bias–variance trade-off similar to JS, but it enjoys additional advantages compared with JS and EB: availability of confidence intervals and flexibility to incorporate covariate information. In our experiments we show that our approach can outperform JS and EB in terms of estimation accuracy (see §5.1).

3 PROBLEM SETTING

Our goal is to assess fairness-related harms of an AI system by evaluating its performance on intersectional subgroups of users specified by $k \geq 2$ sensitive attributes (like race and gender), taking values in finite sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$. The set of all possible k-tuples of sensitive-attribute values is denoted $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_k$.

We assume that we have access to an evaluation dataset S, consisting of individuals described by tuples of the form $(X, A, Y, \hat{Y})$ sampled i.i.d. from some underlying distribution $\mathcal{D}$, where X contains application-relevant information about the individual (e.g., the health history of a patient), $A \in \mathcal{A}$ is a k-tuple of sensitive attributes, Y is an observed outcome variable (e.g., whether the patient was readmitted within 30 days of discharge), and $\hat{Y}$ is an output produced by the AI system (e.g., a score used for prioritizing patients into post-discharge care).

For any k-tuple $a \in \mathcal{A}$, we write $a[1], \ldots, a[k]$ to denote its components. In the ASR example below, we consider two sensitive attributes, *race* and *gender*, with domains $\mathcal{A}_1 = \{\text{Black}, \text{white}\}$ and $\mathcal{A}_2 = \{\text{male}, \text{female}\}$. For $a = (\text{Black}, \text{female})$, we then have $a[1] = \text{Black}$ and $a[2] = \text{female}$. When possible, we use mnemonic indices for components of a and write $a[race]$ and $a[gender]$ to mean $a[1]$ and $a[2]$, and similarly $\mathcal{A}_{race}$ and $\mathcal{A}_{gender}$ to mean $\mathcal{A}_1$ and $\mathcal{A}_2$.

For each $a \in \mathcal{A}$, we define $\mathcal{D}_a$ to be the distribution of individuals with $A = a$, so $\mathcal{D}_a$ is the conditional distribution $\mathcal{D}(X, A, Y, \hat{Y} \mid A = a)$, representing an intersectional group. Let Δ denote the set of all probability distributions over tuples $(X, A, Y, \hat{Y})$, so $\mathcal{D} \in \Delta$ and also $\mathcal{D}_a \in \Delta$ for all $a \in \mathcal{A}$. A *performance metric* is a function $m : \Delta \to \mathbb{R}$ that maps a probability distribution over tuples $(X, A, Y, \hat{Y})$ into a real number. For example, if the underlying AI system performs binary classification, so $Y, \hat{Y} \in \{0, 1\}$, we could measure its performance using *accuracy*, defined, for any $p \in \Delta$, as

$$\text{ACC}(p) = \mathbb{P}_p[Y = \hat{Y}],$$

where $\mathbb{P}_p[\cdot]$ is the probability of an event with respect to p. The overall system performance is then quantified by $\text{ACC}(\mathcal{D})$ and the performance on a group $a \in \mathcal{A}$ by $\text{ACC}(\mathcal{D}_a)$.

Given a performance metric m, the goal of *disaggregated evaluation* is to estimate, for all $a \in \mathcal{A}$, the values

$$\mu_a = m(\mathcal{D}_a).$$

Our only source of information about $\mathcal{D}$ is the evaluation dataset S of size $n = |S|$, sampled i.i.d. from $\mathcal{D}$. The *standard approach* to disaggregated evaluation splits the dataset S into groups $S_a = \{(X, A, Y, \hat{Y}) \in S : A = a\}$ of size $n_a = |S_a|$, and then evaluates m on each S_a (or, more precisely, on the probability distribution that puts an equal probability mass on each data point in S_a). We denote the resulting *standard estimates* by

$$Z_a = m(S_a).$$

For example, if m is accuracy, then

$$Z_a = \text{ACC}(S_a) = \frac{1}{n_a} \sum_{(X, A, Y, \hat{Y}) \in S_a} \mathbf{1}\{Y = \hat{Y}\},$$

where $\mathbf{1}\{\cdot\}$ is an indicator equal to 1 if its argument is true and 0 if it is false.

We next connect this abstract framework to two concrete scenarios already mentioned in §2.

Example 1 (Diabetes). We consider an AI system that refers high-risk patients into a post-discharge care program. We wish to assess allocative harms of this system. To explore this scenario, we use a publicly available dataset of diabetes patients developed by Strack et al. [29]. The dataset contains information about patient hospital visits, including whether each patient was readmitted within 30 days after discharge. We use the readmission as a proxy for whether the patient should be recommended for the care program.

Each data point corresponds to a patient admission, where X describes the patient history and hospital tests; A describes the patient's *race*, *gender*, and (binned) *age*; $Y \in \{0, 1\}$ indicates whether the patient was readmitted within 30 days after discharge; and $\hat{Y} \in [0, 1]$ is the score produced by the AI system that has been trained to predict Y. We assume that the hospital uses a threshold r, and patients with $\hat{Y} \geq r$ are automatically referred into the care program.

One type of allocative harm occurs when a subgroup of patients is disproportionately under-prioritized, i.e., if a subgroup has a low *selection rate*, denoted as

$$\text{SEL}(\mathcal{D}_a) = \mathbb{P}_{\mathcal{D}_a}[\hat{Y} \geq r].$$

We also consider a second type of harm, which occurs when a subgroup of patients experiences a disproportionately large rate of false negatives (i.e., many of those patients that should be recommended are not), measured by the *false negative rate*

$$\text{FNR}(\mathcal{D}_a) = \mathbb{P}_{\mathcal{D}_a}[\hat{Y} < r \mid Y = 1].$$

Example 2 (ASR). To assess quality-of-service harms of an ASR system, we use a dataset from Koenecke et al. [21], consisting of audio snippets (of length between 5s and 50s) spoken by various speakers. In the dataset, X describes properties of the snippet (like duration in seconds), A has two components corresponding to the speaker's *race* and *gender*, Y is the ground-truth transcription of the snippet, and $\hat{Y}$ is the transcription provided by the AI system.

The quality-of-service harms occur when the system underperforms for a subgroup of users. The performance is evaluated by the *word error rate*

$$\text{WER}(\mathcal{D}_a) = \mathbb{E}_{\mathcal{D}_a}[wer(\hat{Y}, Y)],$$

where *wer* is a snippet-level word error rate defined as

$$wer(\hat{Y}, Y) = \frac{subst + del + ins}{|Y|},$$

where *subst*, *del*, and *ins* is the number of word substitutions, deletions, and insertions in $\hat{Y}$ compared with the ground truth Y, and $|Y|$ is the number of words in Y.

To quantify the accuracy of an estimator, like the standard estimator introduced above, we often use *mean squared error* (MSE). We use a modified definition of MSE that accounts for the fact that estimates like $Z_a = m(S_a)$ are sometimes undefined, for instance, when the metric m is defined as a conditional probability, like FNR in Example 1, and the set S_a has no samples that satisfy the condition (e.g., no samples with $Y = 1$ in case of FNR). For an estimator $\hat{\mu}$ of a quantity μ, let $\mathcal{E}$ denote the event that $\hat{\mu}$ is defined. The *bias*,
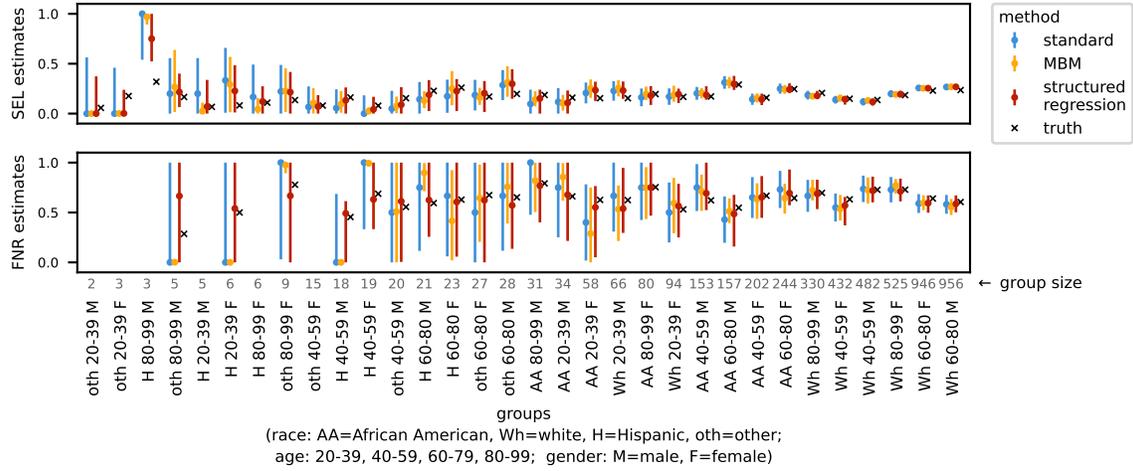
Figure 1: *Point estimates and 95% confidence intervals of selection rate (SEL) and false negative rate (FNR) on diabetes data. Confidence intervals of the standard estimator are calculated using pooled variance (see Eqs. 3 and 4).*

variance, and *mean squared error* (MSE) of $\hat{\mu}$ are defined as

$$\text{Bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu} \mid \mathcal{E}] - \mu,$$

$$\text{Var}(\hat{\mu}) = \mathbb{E}\big[(\hat{\mu} - \mathbb{E}[\hat{\mu} \mid \mathcal{E}])^2 \mid \mathcal{E}\big],$$

$$\text{MSE}(\hat{\mu}) = \mathbb{E}\big[(\hat{\mu} - \mu)^2 \mid \mathcal{E}\big],$$

where the expectations are with respect to the data-generating process giving rise to the dataset used to calculate $\hat{\mu}$ (which is itself a random variable). An estimator with bias equal to zero is called *unbiased*.

Mean squared error decomposes into bias and variance terms as

$$\text{MSE}(\hat{\mu}) = \big[\text{Bias}(\hat{\mu})\big]^2 + \text{Var}(\hat{\mu}),$$

so for unbiased estimators, mean squared error is equal to variance.

Throughout the paper, we assume that the standard estimates Z_a are unbiased. Writing this condition in terms of the metric m, we assume that for all $\mathcal{D} \in \Delta$ and all $n \geq 1$, the performance metric m satisfies

$$\mathbb{E}_{S \sim \mathcal{D}^n}[m(S) \mid m(S) \text{ is defined}] = m(\mathcal{D}), \quad (1)$$

which is true for all the metrics in this paper. Substituting $\mathcal{D}_a$ for $\mathcal{D}$ and n_a for n in Eq. (1) implies that $\mathbb{E}[Z_a \mid Z_a \text{ is defined}] = \mu_a$. In the rest of the paper we drop conditioning on the events like "Z_a is defined," and just write $\mathbb{E}[Z_a] = \mu_a$ for simplicity.

Since the standard estimates Z_a are unbiased, their MSE is equal to their variance, which typically scales as $O(1/n_a)$, the inverse of the number of samples in the group. Thus, standard estimates are accurate when n_a is large, but less so when n_a is small. Unfortunately, even for moderately sized evaluation datasets, the sizes of intersectional groups can be quite small. In Figure 1, we show standard estimates of SEL and FNR on diabetes data (alongside estimates produced by methods introduced later in the paper), along with group sample sizes, n_a (almost half of which are less than 25).

4 STRUCTURED REGRESSION APPROACH

We next develop a structured regression (SR) approach, which seeks to overcome the main shortcoming of the standard estimator: its

large variance for small groups. Our approach builds on two main ideas. First, we *enable variance reduction* by leveraging information across *all data points*, not just data points in S_a, to estimate μ_a. We do this by pooling the data across related groups, for example, across intersectional groups that agree in one of their attributes (like *age*), and by using additional explanatory variables (like X), both of which is accomplished by fitting a regression model for μ_a's, viewing Z_a's as observations thereof. Second, we ensure our regression model is always correctly specified (and thus can produce an unbiased estimate) by including all k-way interaction features. This means our regression model can recover the standard unbiased estimates Z_a as a special case. Regularization is used to optimize the bias–variance trade-off between the high-variance standard estimator and a high-bias (but low-variance) constant estimator.

To start, since the standard estimates are unbiased, i.e., $\mathbb{E}[Z_a] = \mu_a$, we can write

$$Z_a = \mu_a + \varepsilon_a$$

for all $a \in \mathcal{A}$, where ε_a's are independent random variables with $\mathbb{E}[\varepsilon_a] = 0$. We denote the variance of Z_a as $\sigma_a^2 = \text{Var}(Z_a) = \mathbb{E}[\varepsilon_a^2]$. In order to estimate μ_a, we consider a linear model of the form

$$\mu_a = \theta_0 + \boldsymbol{\theta} \cdot \boldsymbol{\phi}^a$$

for all $a \in \mathcal{A}$, where $\boldsymbol{\phi}^a \in \mathbb{R}^d$ is the feature vector describing the group a, and $\theta_0 \in \mathbb{R}$, $\boldsymbol{\theta} \in \mathbb{R}^d$ are the parameters of the linear model. It remains to specify how to define $\boldsymbol{\phi}^a$, how to fit the parameters θ_0 and $\boldsymbol{\theta}$, and how to estimate σ_a.

Defining feature vectors $\boldsymbol{\phi}^a$. The coordinates of $\boldsymbol{\phi}^a$ are referred to as features and denoted as ϕ_j^a for j from some suitable index set. We allow features to be linearly dependent. We consider the following types of features:

(1) *Sensitive features.* These are derived directly from a. We always include group-identity indicators for all the groups $a' \in \mathcal{A}$, yielding features of the form $\phi_{a'}^a = \mathbf{1}\{a = a'\}$. This allows the linear model to express any combination of values μ_a. Additionally, in order to pool information across
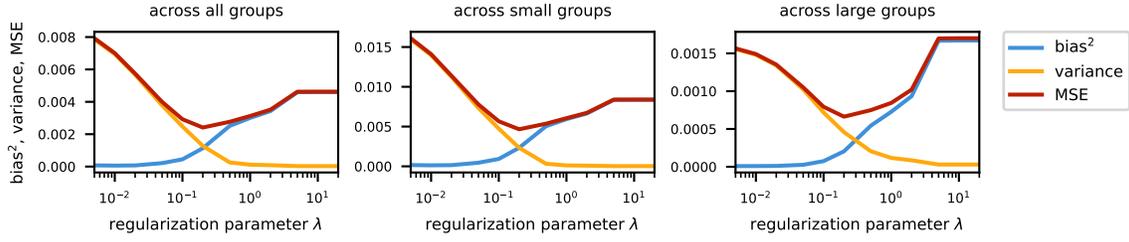
Figure 2: Bias–variance trade-off of structured regression estimates of selection rate (SEL) on diabetes data. Averaged across all groups, small groups (size at most 25), and large groups (size above 25), across 100 draws of evaluation dataset. The scale of the MSE is different for different group sizes, but the minimum MSE is attained around the same value of λ, thanks to the weighting of the training loss.

related groups, we also define indicators for individual attribute values, that is, features of the form $\phi_{i,v}^a = \mathbf{1}\{a[i] = v\}$ for $i \in \{1, \ldots, k\}$ and $v \in \mathcal{A}_i$. In our diabetes example, there are three sensitive attributes: *race*, *age*, and *gender*, with $|\mathcal{A}_{race}| = 4$, $|\mathcal{A}_{age}| = 4$, and $|\mathcal{A}_{gender}| = 2$, so $|\mathcal{A}| = 4 \cdot 4 \cdot 2 = 32$. We use a total of 42 sensitive features: 32 group-identity indicators, 4 indicators of *race*, 4 indicators of *age*, and 2 indicators of *gender*. An example of a group-identity indicator is $\phi_{(\text{Hispanic},80-99,\text{female})}^a$ and an example of a sensitive-attribute indicator is $\phi_{race,\text{Hispanic}}^a$.

(2) *Explanatory features.* These are derived from X, Y, and possibly $\hat{Y}$. We first featurize X using some real-valued functions $f_j(X)$, $j = 1, \ldots, \ell$, and then define explanatory features $\phi_j^a = \mathbb{E}_{X \sim S_a}[f_j(X)]$ (i.e., the average of the feature for group a). Additionally, when Y is categorical, we define features $\phi_y^a = \mathbb{P}_{Y \sim S_a}[Y = y]$ measuring rates of different outcomes in the group a. In our diabetes example, we use 7 explanatory features: 5 are derived from individual-level features f_j, including, for example, the number of inpatient days of a given patient in the prior year; and 2 are of the form ϕ_y^a corresponding to 2 possible values of Y.

(3) *Interaction terms.* It is also possible to consider various interaction terms, both among features of the same type (e.g., between *gender* and *age* indicators), or of different types (e.g., between the outcome Y and *age*).

Fitting the linear model. We fit $(\theta_0, \boldsymbol{\theta})$ by lasso regression [31], minimizing an ℓ_1-penalized square loss. To improve the statistical efficiency of the estimator, loss for each group a is weighted inversely proportional to the variance of Z_a. Intuitively, since our model can express the true μ_a, we expect the square loss on each group to be on the order of the variance of Z_a, so inverse weighting "equalizes the scale" of losses across groups. The penalized loss is then

$$L_\lambda(\theta_0, \boldsymbol{\theta}) = \sum_{a \in \mathcal{A}} \frac{1}{\sigma_a^2} \left(\theta_0 + \boldsymbol{\theta} \cdot \boldsymbol{\phi}^a - Z_a \right)^2 + \lambda \|\boldsymbol{\theta}\|_1, \quad (2)$$

where λ is the regularization hyperparameter. Denoting the minimizer of L_λ (for a given λ) as $(\hat{\theta}_0, \hat{\boldsymbol{\theta}})$, we obtain the *structured regression estimates* $\hat{\mu}_a = \hat{\theta}_0 + \hat{\boldsymbol{\theta}} \cdot \boldsymbol{\phi}^a$.

Tuning λ allows us to navigate the bias–variance trade-off. Because sensitive features include indicators of all values $a \in \mathcal{A}$, when $\lambda = 0$, the loss is minimized by $\hat{\mu}_a = Z_a$ (i.e., we recover the

standard estimates). As $\lambda \to \infty$, the optimization returns $\hat{\boldsymbol{\theta}} \approx \mathbf{0}$. Fixing $\hat{\boldsymbol{\theta}} = \mathbf{0}$ and optimizing only over the intercept term yields the constant solution

$$\hat{\mu}_a = \hat{\mu}_0 \text{ for all } a \in \mathcal{A}, \quad \text{with} \quad \hat{\mu}_0 = \frac{\sum_{a \in \mathcal{A}} Z_a / \sigma_a^2}{\sum_{a \in \mathcal{A}} 1 / \sigma_a^2},$$

corresponding to a weighted average of Z_a's. This solution has a small variance, but it may suffer from a large bias when the true values μ_a are far from identical. By tuning λ, we thus move between the high-variance standard estimate and the high-bias (but low-variance) constant estimate. The mean squared error is typically minimized at some intermediate value of λ (see Figure 2). We tune λ by 10-fold cross-validation, where the individual folds are obtained by stratified sampling of the dataset S with respect to the sensitive attribute tuple A. We can expect our method to be *consistent* (i.e., converge to the true values μ_a as all subgroup sizes n_a grow) because the standard approach is consistent, and is included as the special case of our method with $\lambda = 0$.

Estimating variance σ_a^2. Variances σ_a^2 are needed to calculate weights in our optimization procedure. A simple approach is to estimate σ_a^2 separately on each dataset S_a by using standard variance estimators (when available), or, more generically, by bootstrap. Unfortunately, for small sample sizes, these variance estimates themselves might be inaccurate.

To overcome this limitation, we posit a parametric model for variance, namely, $\sigma_a^2 = \sigma^2 / n_a$, for some parameter σ. To estimate σ, we proceed in two stages. We first use bootstrap on each set S_a to obtain the initial estimate of σ_a^2, which we denote $(\hat{\sigma}_a^{\text{boot}})^2$. Thus, $n_a (\hat{\sigma}_a^{\text{boot}})^2$ is the initial estimate of σ^2. We expect the variance of this estimate to be on the order $O(1/n_a)$. Taking a weighted average across groups, with weighting inversely proportional to $(1/n_a)$, yields our final estimator of σ^2, which translates into an estimator of σ_a^2:

$$\hat{\sigma}^2 = \frac{\sum_{a \in \mathcal{A}} n_a \cdot \left[n_a (\hat{\sigma}_a^{\text{boot}})^2 \right]}{\sum_{a \in \mathcal{A}} n_a},$$

$$\hat{\sigma}_a^2 = \hat{\sigma}^2 / n_a \text{ for all } a \in \mathcal{A}. \quad (3)$$

We refer to these as the *pooled estimates* of variance. In our preliminary experiments, these performed better than the initial estimates $(\hat{\sigma}_a^{\text{boot}})^2$, particularly on small datasets. We note that even if we severely misestimate the variances σ_a^2, our estimation method remains consistent. This is because estimates $\hat{\sigma}_a$ appear only as weights in

the objective (2). Misspecifying the weights will negatively impact the efficiency of the estimator, but not its consistency.

4.1 Confidence intervals

So far we have focused on obtaining point estimates $\hat{\mu}_a$. However, in order for these estimates to be useful in practice, we also need to quantify our uncertainty about their values. We do so by using confidence intervals. For unbiased estimators, like the standard estimator Z_a, confidence intervals can be derived by estimating the variance and then using normal approximation, which works quite well for Z_a with the pooled estimates of variance (see Appendix A).

This approach does not work with lasso estimates, because they are biased—in fact, they achieve their improved accuracy *by being biased*—and so simple confidence intervals derived from variance estimates or bootstrap percentiles are too narrow. Fortunately, there is a rich literature on lasso-based confidence intervals [17, 33, 36]. We use the *residual bootstrap lasso+partial ridge* (rBLPR) approach of Liu et al. [22]. As the name suggests, it is based on a two-stage *lasso+partial ridge* (LPR) point estimator, which first runs lasso as a feature-selection method, and then fits a ridge regression model, which only penalizes the features that were not selected by lasso. The rBLPR method calculates confidence intervals for the LPR estimate by residual bootstrap (see [22] for details).

4.2 Understanding structure of performance variation through goodness-of-fit testing

When presenting the results of disaggregated evaluations, the most common approach is to display point estimates and (sometimes) confidence intervals for every subgroup, as we see, for example, in Figure 1. While this type of a plot can be helpful in identifying groups that may experience poor performance or allocation, it does not provide a narrative for understanding how these harms accrue. Goodness-of-fit testing can complement disaggregated evaluations by allowing us to answer questions such as:

(1) *Do intersectional groups experience additive, sub-additive, or super-additive fairness-related harms?* For example, when a model is found to perform poorly for Black women, is this explained by the model performing poorly for Black people and women, or are there additional sources of error specific to the intersectional group of Black women? An answer to this question can, for example, inform future collection of training data.

(2) *Are there benign factors that explain a significant amount of the observed performance variation across groups?* For example, are observed differences in the performance of an ASR system attributable to systematically worse audio quality in the recordings for speakers from certain groups? Presence of such benign factors *does not lessen the harm*, but the knowledge of the factors that drive performance differences can be used to design mitigations (for example, denoising algorithms targeted at specific types of sensors or noise characteristics).

These types of questions can be framed as goodness-of-fit tests. We consider goodness-of-fit tests that compare two linear models: M_0, with fewer features, and M_1, with some additional features. Such a test asks whether the additional features included in model M_1 improve the goodness of fit compared with model M_0, where the goodness of fit is measured using the square loss as in Eq. (2). To answer the first question above, we can compare a model M_0, which includes only indicators of *race* and *gender*, with a model M_1, which also includes interaction terms. To answer the second question, we can compare a model M_0', which only includes benign factors, with a model M_1', which additionally includes indicators of *race*, *gender*, and *age*.

While there are goodness-of-fit tests that have been designed for lasso regression [16], in this paper, we use standard F-tests designed for unregularized linear regression. In contrast to the foregoing discussion, we do not include features corresponding to the indicators of a (because these would trivially yield the standard estimates Z_a with 0 residual sum-of-squares (RSS), which in this case corresponds to overfitting).

5 EXPERIMENTS

In this section, we evaluate the accuracy of point estimates and calibration of confidence intervals produced by our structured regression (SR) approach. We also demonstrate how goodness-of-fit tests can be used to provide insights about what drives the variation of performance across groups.

We compare SR with several baselines. First, there is the **standard estimator** $Z_a = m(S_a)$. We construct confidence intervals for Z_a using normal approximation with pooled variance estimates $(\hat{\sigma}_a)^2$ from Eq. (3). Given a confidence level γ (say 95%), or a significance level $\alpha = 1 - \gamma$ (say 5%), we use the confidence interval

$$[Z_a + q_{\alpha/2}\hat{\sigma}_a, \ Z_a + q_{1-\alpha/2}\hat{\sigma}_a], \tag{4}$$

where q_p is the p-th quantile of the standard normal distribution.

Our second baseline is the **model-based metrics (MBM)** approach [23]. As mentioned in §2, MBM is a Bayesian approach to structured regression that models the scores produced by an AI system (like $\hat{Y}$ in the diabetes example). It is not directly applicable to performance metrics that are not based on scores, so we do not use it in the ASR experiments. Similar to SR, MBM uses linear modeling, and so requires specifying features for each data point. It comes with a boostrapping procedure for constructing confidence intervals.

We also compare our point estimates with the classical **James-Stein (JS)** estimator [15, 28]. The estimator works by shrinking standard estimates towards zero (or some other constant). We use a variant due to Bock [4], which is adapted to unequal variances (in our case, pooled estimates $\hat{\sigma}_a^2 = \hat{\sigma}^2/n_a$), giving rise to

$$\hat{\mu}_a^{JS} = \hat{\mu}_0 + \left(1 - \frac{(|\mathcal{A}| - 3)\hat{\sigma}^2}{\sum_{a' \in \mathcal{A}} n_{a'}(Z_{a'} - \hat{\mu}_0)^2}\right)_+ (Z_a - \hat{\mu}_0),$$

where $\hat{\mu}_0 = (\sum_{a \in \mathcal{A}} n_a Z_a)/n$ is a weighted average of Z_a's. Compared with Bock's original estimator [4], we use $|\mathcal{A}|$ in the numerator, as this has been previously observed to lead to better performance [9]. Since $\hat{\mu}^{JS}$ is not an unbiased estimator, construction of confidence intervals presents a challenge and we are not aware of any standard procedure.

And finally, we compare our method with the **empirical Bayes (EB)** approach [8], which posits a hierarchical Bayesian model, and then estimates μ_a by posterior means, while fixing hyperparameters to their point estimates. In Appendix B we derive the following
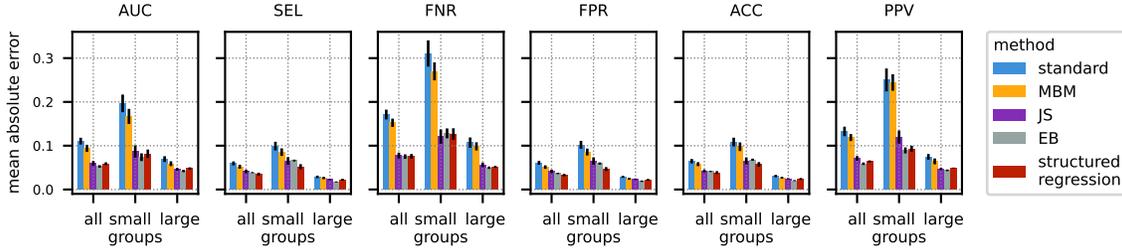
Figure 3: *Mean absolute error of estimates of 6 metrics using 5 methods on diabetes data.* **Averaged across all groups, small groups (size at most 25), and large groups (size above 25), across 20 draws of evaluation dataset.**

variant, which we use in our experiments:

$$\hat{\mu}_a^{\text{EB}} = \hat{\mu} + \left(1 - \frac{\hat{\sigma}_a^2}{\hat{\tau}^2 + \hat{\sigma}_a^2}\right)(Z_a - \hat{\mu}),$$

where $\hat{\sigma}_a^2$ is the pooled estimate of variance, and $\hat{\tau}^2$ and $\hat{\mu}$ are obtained by

$$\hat{\tau}^2 = \left(\frac{\sum_{a \in \mathcal{A}} n_a (Z_a - \hat{\mu}_0)^2 - (|\mathcal{A}| - 1)\hat{\sigma}^2}{n - \sum_a n_a^2/n}\right)_+,$$

$$\hat{\mu} = \frac{\sum_{a \in \mathcal{A}} Z_a/(\hat{\tau}^2 + \hat{\sigma}_a^2)}{\sum_{a \in \mathcal{A}} 1/(\hat{\tau}^2 + \hat{\sigma}_a^2)}.$$

Similar to JS, we are not aware of any standard procedure for construction of confidence intervals.

5.1 Diabetes experiments

We first explore the scenario from Example 1 using the dataset developed by Strack et al. [29], and previously used in an AI fairness tutorial [13] and to evaluate the MBM approach [23]. The dataset contains hospital admission records from 130 hospitals in the U.S. over a ten-year period (1998–2008) for patients who were admitted with a diabetes diagnosis and whose hospital stay lasted one to fourteen days. It is a tabular dataset with 47 features describing each encounter, including patient demographics and clinical information. Following Miller et al. [23], we filter out records with missing demographics and those with age below 20. We preprocess clinical features as in [13]. To emulate an AI system that scores patients for a post-discharge care program, we use 25% of the data to train a logistic regression model to predict whether the patient will be readmitted into hospital within 30 days. The remaining 75% of the data, consisting of 73,988 hospital admissions across 55,157 individuals, is used as the ground truth $\mathcal{D}$ in all of our evaluation experiments.

We consider three sensitive attributes, *race*, *age*, and *gender*, with $\mathcal{A}_{race} = \{$African American, Hispanic, white, other$\}$, $\mathcal{A}_{age} = \{20$–$39, 40$–$59, 60$–$79, 80$–$99\}$, and $\mathcal{A}_{gender} = \{$male, female$\}$. Hospital admissions are represented as tuples $(X, A, Y, \hat{Y})$, where X contains clinical features, $A = (race, age, gender)$, $Y \in \{0, 1\}$ indicates whether the patient was readmitted within 30 days of discharge, and $\hat{Y} \in [0, 1]$ is the readmission probability predicted by the logistic regression model. From the ground truth we then sample an evaluation dataset S of size 5000 by stratified sampling according to A.

As in Example 1, we assume that the hospital uses a threshold r, and patients with $\hat{Y} \geq r$ are automatically referred into the care program. We set the threshold r so that $\mathbb{P}_{\mathcal{D}}[\hat{Y} \geq r] = 0.2$, meaning

that only 20% of patients are referred, and write $\pi(\hat{Y}) = \mathbf{1}\{\hat{Y} \geq r\}$ to denote this decision rule. We consider six performance metrics (including those already introduced earlier), defined for any $p \in \Delta$ as

$$\text{SEL}(p) = \mathbb{P}_p[\pi(\hat{Y}) = 1],$$
$$\text{ACC}(p) = \mathbb{P}_p[\pi(\hat{Y}) = Y],$$
$$\text{FNR}(p) = \mathbb{P}_p[\pi(\hat{Y}) = 0 \mid Y = 1],$$
$$\text{FPR}(p) = \mathbb{P}_p[\pi(\hat{Y}) = 1 \mid Y = 0],$$
$$\text{PPV}(p) = \mathbb{P}_p[Y = 1 \mid \pi(\hat{Y}) = 1],$$
$$\text{AUC}(p) = \mathbb{P}_{(Y,\hat{Y}) \sim p, (Y',\hat{Y}') \sim p}[\hat{Y} < \hat{Y}' \mid Y = 0, Y' = 1].$$

The first five metrics (selection rate, accuracy, false positive rate, false negative rate, and positive predictive value) are derived from the confusion matrix. The final metric is the area under the ROC curve; $(Y, \hat{Y})$ and $(Y', \hat{Y}')$ in its definition are sampled independently according to p.

In order to apply SR, we need to specify features ϕ^a. As sensitive features, we use indicators of *race*, *age*, *gender*, as well as indicators of the triple (*race*,*age*,*gender*). We use 7 explanatory features: indicators for 2 possible values of Y, and 5 additional clinical features describing the number of inpatient visits, outpatient visits, and emergency visits in the preceding year, number of diagnoses at admission, and whether any of the diagnoses was congestive heart failure. For MBM, we use the same set of features, but without the triple indicators.

In Figure 1 from earlier, point estimates obtained by SR appear to be closer to the ground truth than those obtained by the standard method and MBM. Confidence intervals constructed by SR are of similar size as the standard confidence intervals, and occasionally smaller. MBM appears to produce smaller confidence intervals than SR, but they seem to miss the ground-truth metric values more often. We next evaluate these anecdotal observations more systematically.

In Figure 3, we evaluate the quality of point estimates using *mean absolute error* (MAE), which is the mean deviation of the point estimate from the truth, averaged across 20 draws of evaluation dataset, and over all groups, or separately over the groups of size at most 25 (which we call *small*) and groups of size greater than 25 (which we call *large*). JS, EB and SR yield substantially more accurate point estimates than the standard method and MBM. The improvement is particularly dramatic for small groups. JS, EB and SR all perform at a similar level, but SR tends to work best on small groups, and EB is marginally better than JS and SR on large groups (see Figure 7 in Appendix D for a comparison limited to these three

Christine Herlihy, Kimberly Truong, Alexandra Chouldechova, and Miroslav Dudík
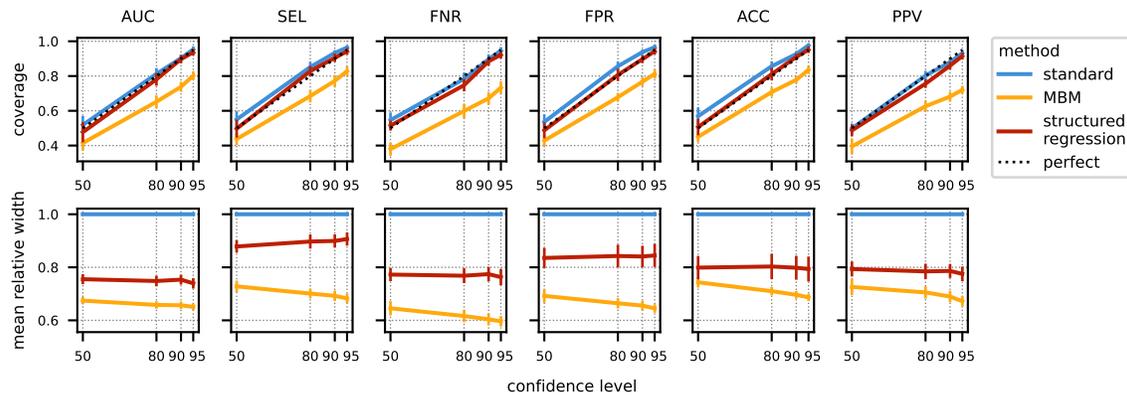
Figure 4: *Coverage and mean relative width of confidence intervals for 6 metrics constructed by 3 methods on diabetes data.* Averaged across all groups and across 20 draws of evaluation dataset. Relative width is with respect to the width of the standard confidence interval.

Table 1: *Goodness-of-fit tests on diabetes data.* From left to right, we consider increasingly more complex models with a growing set of features and report the p-values of the associated goodness-of-fit tests; p-values below 0.05 are in bold.

Estimated metric	Goodness-of-fit test p-values (comparing a more expressive *vs* a less expressive model)			
	expl vs $\emptyset$	*sens vs* $\emptyset$	*expl + sens vs expl*	*expl + sens + Y · sens vs expl + sens*
AUC	*0.281*	*0.438*	*0.726*	*0.543*
SEL	**0.000**	**0.000**	**0.000**	*0.842*
FNR	*0.093*	*0.473*	*0.735*	*0.565*
FPR	**0.001**	**0.000**	**0.000**	*0.431*
ACC	**0.000**	**0.000**	**0.005**	*0.182*
PPV	*0.316*	*0.493*	*0.470*	*0.874*

Model abbreviations: $\emptyset$=intercept only, expl=explanatory features, sens=sensitive features, ·=interactions

methods). We use MAE instead of MSE, because MAE values are easier to interpret, but MSE results are qualitatively similar.

In Figure 4, we shift attention to confidence intervals. In the top plots, we evaluate *coverage*, that is, how often the ground truth lies in the confidence intervals (across 20 draws of evaluation dataset and across all groups). We show coverage as a function of the confidence level. We see that both standard method as well as SR are well-calibrated, with their coverage close to the confidence level, whereas MBM is over-confident, with coverage well below the confidence level. In the bottom plots, we evaluate the *mean relative width* of confidence intervals, meaning the mean of the ratio between the width of a confidence interval and the width of the standard confidence interval. We see that MBM has the narrowest intervals, but this is at the expense of coverage. On the other hand, SR is able to maintain well-calibrated coverage while still decreasing the confidence intervals by up to 20% compared with the standard method.

Finally, in Table 1, we demonstrate the use of goodness-of-fit tests. We begin with the question: *Is there statistically significant*

evidence of performance disparity across groups; and if so, is there further evidence of intersectional harm? Table 1 shows p-values for goodness-of-fit tests beginning with just the intercept, adding explanatory features, then sensitive features (just the indicators of *race*, *age*, and *gender*, but not of their combination), and eventually interaction terms between the outcome Y and sensitive features. There is no evidence to go beyond the intercept-only model when estimating AUC, FNR, PPV. That is, there is no *detectable* variation in AUC, FNR, or PPV across groups (or other explanatory variables). For FNR, this is consistent with what we observe in Figure 1. The confidence intervals shown are large and overlapping for the vast majority of groups, even after SR is applied to help reduce uncertainty. Because of how wide the FNR confidence intervals are, the reasonable conclusion is that sample sizes are too small for the inference to be conclusive, and *not* that we have definitive evidence of equal performance across groups. On the other hand, the table shows that both explanatory and sensitive features help with modeling SEL, FPR, and ACC. In fact, sensitive features improve the fit *after* the explanatory features have already been added, meaning that differences in performance across the groups cannot be explained by the "benign" explanatory features alone.

In Appendix C we provide more examples of insight that may be gained through goodness-of-fit analysis using semi-synthetic variants of the diabetes data generated to exhibit different ground-truth structures for the underlying variation in system performance.

5.2 Experiments with ASR data

We now explore the scenario from Example 2 using the data provided by Koenecke et al. [21] as a supplement to their paper finding racial disparities in commercial ASR systems. Similar to Koenecke et al. [21], we use the *matched* dataset, which contains 4282 snippets across 105 distinct speakers. (Matching ensures that there is the same number of snippets from Black and white speakers and that the marginal distributions of various descriptive statistics match.)

For each audio snippet, we are provided with various statistics (like duration and word count), an anonymized speaker id, speaker demographics, and word error rates (WERs) on that snippet by five ASR systems (by Google, IBM, Amazon, Microsoft, and Apple).
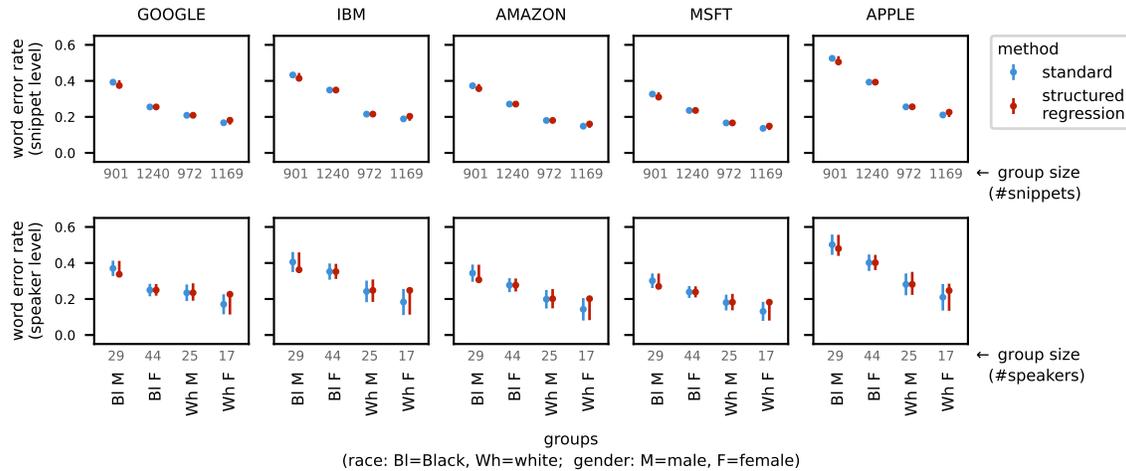
Figure 5: *Point estimates and 95% confidence intervals of word error rates of five ASR systems.*

This information is encoded as a tuple $(X, A, W_1, \ldots, W_5)$, where X contains the identity of the speaker, the duration of the snippet in seconds, and word count, A contains two sensitive attributes, *gender* and *race*, $\mathcal{A}_{gender} = \{\text{male}, \text{female}\}$ and $\mathcal{A}_{race} = \{\text{Black}, \text{white}\}$, and finally, instead of Y (human transcription) and $\hat{Y}_1, \ldots, \hat{Y}_5$ (transcriptions by five ASR systems), we directly have the corresponding word error rates $W_i = wer(\hat{Y}_i, Y)$. The performance metric for the system i is thus $m(p) = \mathbb{E}_p[W_i]$ for any $p \in \Delta$.

Although there appears to be a large number of samples ($n = 4282$), there are only 105 distinct speakers. We expect there to be a substantial amount of correlation between WERs of the same individual, so an analysis that treats the WERs as independent is likely to overstate the statistical significance of findings, and may arrive at incorrect conclusions, in particular, when some speakers have many more snippets than others. In our experiments, we therefore present results both from a snippet-level analysis that treats the WERs across all snippets as independent (as done in [21]), and a speaker-level analysis that first reduces the data to speaker-level WERs by taking an average of WERs across the speaker's snippets.

We first compare disaggregated evaluation results obtained by SR versus the standard method. To apply SR, we need to specify features ϕ^a. As sensitive features, we use indicators of *race* and *gender*, as well as indicators of the pair (*race*,*gender*). We use only one explanatory feature, equal to the log duration of the snippet.

In Figure 5, we report the results. At the snippet level, both methods generally replicate the results of Koenecke et al. [21]: Black male speakers have the largest WER, followed by Black female speakers, white male speakers, and white female speakers. The main difference is that SR systematically shrinks the WER values of the extreme groups (Black male speakers and white female speakers) towards the mean. Results at the speaker level have substantially larger confidence intervals than the snippet-level results, reflecting smaller group sizes. Also, due to smaller group sizes, the SR point estimates are shrunk towards the mean more aggressively.

We also carry out the goodness-of-fit analysis of structure of intersectional harms. At the speaker level, we find that the variation of performance of all systems is well-explained by the additive

model *expl* + *race* + *gender* (the p-values of adding each variable in turn are below 0.003), but not by a model with interactions. This is in contrast with the snippet-level analysis, which supports the model with interactions (with p-values below 0.001). We interpret this conservatively and conclude that there is evidence for an additive structure of intersectional harms, but not for an interaction term. This does not mean that there are no interaction effects, just that we cannot conclude that from the data at hand.

6 CONCLUSION

We have introduced a structured regression approach to disaggregated evaluation and compared its performance with a variety of baselines. We have seen that the structured regression (SR), James-Stein (JS) and empirical Bayes (EB) estimators all substantially improve accuracy of point estimates compared with the standard approach as well as a more sophisticated MBM baseline. SR, JS and EB are simple to implement, and are also close in terms of performance, so the choice among them should be driven by their usability. Here, SR has some advantages. Its ability to include application-specific features makes it more flexible, and it has a well-developed inference procedures like construction of confidence intervals and goodness-of-fit tests. Examining JS and EB more closely from inference perspective in the context of disaggregated evaluation is a promising direction for future research and a necessity for their practical use. Note that we have evaluated SR only in two domains, so any applications in domains with different characteristics (like the number and types of explanatory and sensitive features, or dataset size) require additional validation.

Many important challenges lie outside the scope of this paper. For example, we assume that relevant sensitive attributes and performance metrics have been determined. However, as Barocas et al. [3] discuss, the sensitive attributes often include socially constructed—and potentially contested—features (like race and gender), which makes the task of mapping people to attributes and corresponding subgroups potentially fraught, particularly when it involves inference or use of proxy variables, or poses a risk for members of already-marginalized subgroups. As a separate challenge, in many

high-stakes applications (like education and healthcare), we are not able to directly measure who might benefit, so we need to rely on proxies. A poor choice of a proxy may further exacerbate existing inequities, as is the case, for instance, when predicting risk of re-offense from arrest records [10] or predicting healthcare needs based on healthcare expenditures [25].

Once the disaggregated results are produced, a complementary set of challenges arises in how to interpret them. We have conspicuously omitted analysis of regression coefficients, because in our preliminary experiments, we found that lasso coefficients exhibit too much variance for reliable inference. Instead, we suggest to use goodness-of-fit tests and have demonstrated several ways how. We acknowledge that we have just taken some initial steps in this area, and there are many opportunities to apply more sophisticated statistical techniques. Our exploration also completely leaves out important sociotechnical questions about how to draw actionable conclusions, and how to best communicate the results to relevant stake-holders, both of which are key in translating fairness assessments into a reduction in fairness-related harms.

ACKNOWLEDGMENTS

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*.

[3] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In *AIES 2021*. ACM.

[4] M. E. Bock. 1975. Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics* 3, 1 (1975).

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91.

[6] Alexandra Chouldechova, Siqi Deng, Yongxin Wang, Wei Xia, and Pietro Perona. 2022. Unsupervised and semi-supervised bias benchmarking in face recognition. In *European Conference on Computer Vision*. Springer, 289–306.

[7] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. Vol. 1989, Article 8.

[8] Bradley Efron and Carl Morris. [n. d.]. Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach. *J. Amer. Statist. Assoc.* 68, 341 ([n. d.]), 117–130.

[9] Sergey Feldman, Maya Gupta, and Bela Frigyik. 2012. Multi-Task Averaging. In *Advances in Neural Information Processing Systems*, Vol. 25.

[10] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. 2020. Fairness evaluation in presence of biased noisy labels. In *International conference on artificial intelligence and statistics*. PMLR, 2325–2336.

[11] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 424–432.

[12] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.

[13] Triveni Gandhi, Manojit Nandi, Miroslav Dudík, Hanna Wallach, Michael Madaio, Hilde Weerts, Adrin Jalali, and Lisa Ibañez. 2021. Fairness in AI Systems: From Social Context to Practice using Fairlearn. Tutorial presented at the 20th annual

Scientific Computing with Python Conference (Scipy 2021), Virtual Event. https://github.com/fairlearn/talks/tree/main/2021_scipy_tutorial

[14] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. PMLR, 1939–1948.

[15] W. James and C. Stein. 1961. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*. 361––379.

[16] Jana Janková, Rajen D Shah, Peter Bühlmann, and Richard J Samworth. 2020. Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 3 (2020), 773–795.

[17] Adel Javanmard and Andrea Montanari. 2014. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *Journal of Machine Learning Research* 15, 82 (2014), 2869–2909.

[18] Disi Ji, Robert L. Logan, Padhraic Smyth, and Mark Steyvers. 2020. Active Bayesian Assessment for Black-Box Classifiers. https://doi.org/10.48550/ARXIV.2002.06532

[19] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. https://doi.org/10.48550/ARXIV.2010.09851

[20] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*. PMLR, 2564–2572.

[21] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117

[22] Hanzhong Liu, Xin Xu, and Jingyi Jessica Li. 2020. *Statistica Sinica* 30, 3 (2020), 1333–1355.

[23] Andrew C. Miller, Leon A. Gatys, Joseph Futoma, and Emily Fox. 2021. Model-based metrics: Sample-efficient estimates of predictive model subpopulation performance. In *Proceedings of the 6th Machine Learning for Healthcare Conference*. PMLR, 308–336. https://proceedings.mlr.press/v149/miller21a.html

[24] Mathieu Molina and Patrick Loiseau. 2022. Bounding and approximating intersectional fairness through marginal fairness. *Advances in Neural Information Processing Systems* 35 (2022), 16796–16807.

[25] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342 arXiv:https://www.science.org/doi/pdf/10.1126/science.aax2342

[26] Vihari Piratla, Soumen Chakrabarti, and Sunita Sarawagi. 2021. Active Assessment of Prediction Services as Accuracy Surface Over Attribute Combinations. *CoRR* abs/2108.06514 (2021). arXiv:2108.06514 https://arxiv.org/abs/2108.06514

[27] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[28] C. Stein. 1956. Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*. 197–206.

[29] Beata Strack, Jonathan Deshazo, Chris Gennings, Juan Luis Olmo Ortiz, Sebastian Ventura, Krzysztof Cios, and John Clore. 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed research international* 2014 (04 2014), 781670. https://doi.org/10.1155/2014/781670

[30] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising. *Queue* 11, 3 (mar 2013), 10–29. https://doi.org/10.1145/2460276.2460278

[31] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 1 (1996), 267–288.

[32] Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. 2023. Can You Rely on Your Model Evaluation? Improving Model Evaluation with Synthetic Test Data. *arXiv preprint arXiv:2310.16524* (2023).

[33] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42, 3 (2014), 1166 – 1202.

[34] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 336–349.

[35] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research* 24, 257 (2023), 1–8. http://jmlr.org/papers/v24/23-0389.html

[36] Cun-Hui Zhang and Stephanie S. Zhang. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76, 1 (2014), 217–242.
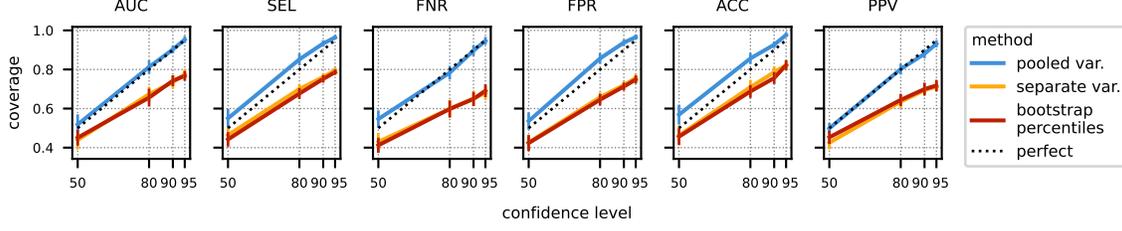
Figure 6: *Comparison of methods for constructing confidence intervals for the standard estimator.* **Showing coverage of confidence intervals constructed for six metrics on diabetes data, averaged over all groups and over 20 draws of evaluation dataset. Confidence intervals constructed from pooled variance are close to the perfect line (corresponding to coverage equal to confidence level). Confidence intervals derived from separately estimated variances undercover true values.**

A CONFIDENCE INTERVALS FOR STANDARD ESTIMATES

We consider three methods for constructing confidence intervals for standard estimates Z_a at a given confidence level γ (e.g., 95%), or equivalently, at a significance level $\alpha = 1 - \gamma$ (e.g., 5%).

Two of the methods are based on normal approximation and take form

$$[Z_a + q_{\alpha/2}\hat{\sigma}_a, \ Z_a + q_{1-\alpha/2}\hat{\sigma}_a],$$

where q_p is the p-th quantile of the standard normal distribution and $\hat{\sigma}_a^2$ is an estimate of variance of Z_a. We consider either the pooled estimate of variance derived in Eq. (3), or the estimate $(\hat{\sigma}_a^{\text{boot}})^2$ obtained by boostrap on S_a. The third method uses bootstrap percentiles on S_a.

In Figure 6, we compare coverage properties of the resulting confidence intervals on diabetes data. Confidence intervals constructed from pooled variance estimates are well-calibrated, with coverage closely matching their confidence level. The other two methods substantially undercover true values.

B DERIVATION OF THE EMPIRICAL BAYES ESTIMATOR

We posit the following hierarchical Gaussian model:

$$\begin{aligned} \mu_a &\sim \mathcal{N}(\mu, \tau^2) \quad \text{for all } a \in \mathcal{A}, \\ Z_a &\sim \mathcal{N}(\mu_a, \sigma_a^2) \quad \text{for all } a \in \mathcal{A}, \end{aligned} \quad (5)$$

where σ_a is known and μ and τ are unknown hyperparameters. We observe values Z_a and need to predict μ_a.

Conditioning on the prior and observations, we obtain the posterior distribution

$$\mu_a \mid \mu, \tau, Z_a \sim \mathcal{N}\left(\hat{\mu}_a^{\text{EB}}, (\hat{\sigma}_a^{\text{EB}})^2\right)$$

where the posterior mean and variance are equal to

$$\hat{\mu}_a^{\text{EB}} = \mu + \frac{\tau^2}{\tau^2 + \sigma_a^2} \cdot (Z_a - \mu), \quad (\hat{\sigma}_a^{\text{EB}})^2 = \frac{\tau^2}{\tau^2 + \sigma_a^2} \cdot \sigma_a^2. \quad (6)$$

The empirical Bayes approach takes point estimates of the hyperparameters μ and τ, and plugs them into Eq. (6). The resulting $\hat{\mu}_a^{\text{EB}}$ is used as a point estimate for μ_a and the resulting $\hat{\sigma}_a^{\text{EB}}$ is used to construct credible intervals for μ_a.

We estimate τ^2 by analyzing a suitable sum of squares (similarly as in the analysis of variance). To start, note that if we marginalize out μ_a from Eq. (5), we find that the values Z_a are conditionally independent given μ and τ, with

$$Z_a \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^2 + \sigma_a^2) \quad \text{for all } a \in \mathcal{A}. \quad (7)$$

For each a, we consider the square $(Z_a - \hat{\mu}_0)^2$, where

$$\hat{\mu}_0 = \sum_{a \in \mathcal{A}} w_a Z_a, \quad \text{with} \quad w_a = n_a/n \text{ for all } a \in \mathcal{A}.$$

The expectation of $(Z_a - \hat{\mu}_0)^2$ then takes the following form:

$$\begin{aligned} &\mathbb{E}\left[\left(Z_a - \hat{\mu}_0\right)^2 \middle| \mu, \tau\right] \\ &= \mathbb{E}\left[\left(\sum_{a' \neq a} w_{a'} Z_{a'} - (1 - w_a) Z_a\right)^2 \middle| \mu, \tau\right] \\ &= \mathbb{E}\left[\left(\sum_{a' \neq a} w_{a'} (Z_{a'} - \mu) - (1 - w_a)(Z_a - \mu)\right)^2 \middle| \mu, \tau\right] \\ &= \sum_{a' \neq a} w_{a'}^2 (\tau^2 + \sigma_{a'}^2) + (1 - w_a)^2 (\tau^2 + \sigma_a^2) \quad (8) \\ &= \sum_{a' \in \mathcal{A}} w_{a'}^2 (\tau^2 + \sigma_{a'}^2) + (1 - 2w_a)(\tau^2 + \sigma_a^2), \quad (9) \end{aligned}$$

where Eq. (8) follows by Eq. (7) and conditional independence of Z_a's. Multiplying Eq. (9) by w_a and summing across all a, we obtain

$$\begin{aligned} &\mathbb{E}\left[\sum_{a \in \mathcal{A}} w_a \left(Z_a - \hat{\mu}_0\right)^2 \middle| \mu, \tau\right] \\ &= \sum_{a' \in \mathcal{A}} w_{a'}^2 (\tau^2 + \sigma_{a'}^2) + \sum_{a \in \mathcal{A}} w_a (1 - 2w_a)(\tau^2 + \sigma_a^2) \\ &= \sum_{a \in \mathcal{A}} w_a (1 - w_a)(\tau^2 + \sigma_a^2). \end{aligned}$$

We rearrange the final expression to obtain an unbiased estimate of τ^2:

$$\hat{\tau}^2 = \frac{\sum_{a \in \mathcal{A}} w_a (Z_a - \hat{\mu}_0)^2 - \sum_{a \in \mathcal{A}} w_a (1 - w_a)\sigma_a^2}{1 - \sum_{a \in \mathcal{A}} w_a^2}. \quad (10)$$

Since $\mathbb{E}[Z_a \mid \mu, \tau] = \mu$ and $\text{Var}[Z_a \mid \mu, \tau] = \tau^2 + \sigma_a^2$, we estimate μ by taking a weighted average of Z_a's, with the weights proportional to the inverse of the variance, but with $\hat{\tau}^2$ plugged in for τ^2:

$$\hat{\mu} = \frac{\sum_{a \in \mathcal{A}} Z_a / (\hat{\tau}^2 + \sigma_a^2)}{\sum_{a \in \mathcal{A}} 1 / (\hat{\tau}^2 + \sigma_a^2)}. \quad (11)$$

Table 2: *Semi-synthetic data generation.* **We consider four semi-synthetic ground-truth structures, which we refer to as model$_{age}$, model$_{age+rc}$, model$_{age\cdot rc}$, and model$_{expl}$, depending which variables they depend on and how, with "+" denoting additive dependence and "·" presence of interaction. Model coefficients have been chosen to ensure that in all cases $\mathbb{E}_{\mathcal{D}}[\hat{Y}] \approx 0.27$ and $(\mathbf{Var}_{\mathcal{D}}[\hat{Y}])^{1/2} \approx 0.44$.**

Model name	Ground-truth value of μ_a	Data-generating process
model$_{age}$	$\mu_a = 0.35 - 0.3 \cdot \phi^a_{age,40-60}$	$\hat{Y} = \text{Bernoulli}(\mu_A)$
model$_{expl}$	$\mu_a = -0.93 + 0.16 \cdot \mathbb{E}_{\mathcal{D}_a}[X_{number\_diagnoses}]$	$\hat{Y} = \mathcal{N}(\mu_A, 0.1)$
model$_{age+rc}$	$\mu_a = 0.65 - 0.15 \cdot \phi^a_{age,40-60} - 0.45 \cdot \phi^a_{race,white}$	$\hat{Y} = \text{Bernoulli}(\mu_A)$
model$_{age\cdot rc}$	$\mu_a = 0.32 - 0.27 \cdot \phi^a_{age,40-60} \cdot \phi^a_{race,white}$	$\hat{Y} = \text{Bernoulli}(\mu_A)$

Table 3: *Goodness-of-fit tests on semi-synthetic data.* **From left to right, we consider increasingly more complex models with a growing set of features and report the p-values of the associated goodness-of-fit tests; p-values below 0.05 are in bold.**

Data-generating model	Goodness-of-fit test p-values (comparing a more expressive *vs* a less expressive model)						
	expl vs ∅	*age vs ∅*	*expl + age vs expl*	*expl + rc vs expl*	*expl + age + rc vs expl + age*	*expl + age + rc vs expl + rc*	*expl + age + rc + age · rc vs expl + age + rc*
model$_{age}$	**0.025**	**0.000**	**0.000**	*0.487*	*0.153*	**0.000**	*0.576*
model$_{expl}$	**0.000**	**0.000**	*0.323*	*0.366*	*0.608*	*0.551*	**0.000**
model$_{age+rc}$	**0.013**	*0.661*	*0.142*	**0.000**	**0.000**	**0.000**	*0.089*
model$_{age\cdot rc}$	**0.003**	**0.000**	**0.000**	**0.040**	**0.015**	**0.000**	**0.002**

Model abbreviations: ∅=intercept only, expl=explanatory features, rc=race, ·=interactions

The last missing piece is σ_a^2, for which we use the pooled estimate from Eq. (3).

Combining it all together, we use the pooled estimates of variance $\hat{\sigma}_a^2$ and the weighted mean $\hat{\mu}_0$ alongside observations Z_a to calculate $\hat{\tau}^2$ using Eq. (10); then we calculate $\hat{\mu}$ using Eq. (11); and finally we calculate $\hat{\mu}_a^{\text{EB}}$ and $(\hat{\sigma}_a^{\text{EB}})^2$ using Eq. (6).

C GOODNESS-OF-FIT EXPERIMENTS WITH SEMI-SYNTHETIC DATA

In this appendix we provide further examples of the kinds of insight concerning the structure of intersectional harm that may be obtained through goodness-of-fit analysis. Our examples rely on semi-synthetic data. More precisely, we continue to use the diabetes dataset as described in §5.1, but with different values $\hat{Y}$. We consider the performance metric $m(p) = \mathbb{E}_p[\hat{Y}]$ (this is quite similar to selection rate or word error rate) and generate $\hat{Y}$ in such a way that ground-truth metric values μ_a have a specific structure.

We consider four semi-synthetic ground-truth structures, which we refer to as model$_{age}$, model$_{age+rc}$, model$_{age\cdot rc}$, and model$_{expl}$, depending which variables they depend on and how, with "+" denoting additive dependence and "·" presence of interactions. These models correspond, respectively, to settings where the true variation in model performance is explained entirely by *age* alone; by *age* and *race* additively, with middle-age patients and white patients experiencing lower values of the performance metric; by *age* and *race* interactionally, with the intersectional group of *white middle-age* patients experiencing lower values of the performance metric whereas middle-age patients of other races and white patients of other ages do not; and entirely by *non-demographic explanatory*

factors. The details of the underlying data generation process are summarized in Table 2.

Table 3 summarizes the results of a sequence of goodness-of-fit tests set up to investigate different questions about the structure of the underlying performance disparity using the observed semi-synthetic data. Moving from left to right, we test goodness-of-fit of more and more complex models. The first column, *expl vs ∅* asks: *is the data consistent with constant model performance that does vary with available (non-demographic) explanatory variables?* In the first row, the ground-truth depends only on *age*, but we see a statistically significant improvement in goodness-of-fit from ∅ to *expl*, because of correlation between *age* and *expl*. The third column then asks: *is the variation in performance explained by the benign explanatory factors alone, or is there evidence of further variation with age?* After the explanatory features have been included, *age* still helps (the improvement from *expl* to *expl + age* is significant), so the variation in the performance metric cannot be explained by the "benign explanatory factors" alone. While not presented here, the results of *expl + age vs age* are *not* statistically significant in this setting, confirming that the benign factors do not explain additional variation that is not explained by age alone.

Shown in the second row is the setting where the data is drawn from model$_{expl}$. Columns 3 and 4 ask whether there is additional variation in performance in *age* or *race*, respectively, that is not explained by the benign factors. The p-values for these goodness-of-fit tests are not statistically significant, so we conclude correctly that there is no evidence that *age* or *race* help after explanatory features have been added.

In the last two rows of Table 3, we demonstrate how goodness-of-fit tests handle data from additive models versus models with
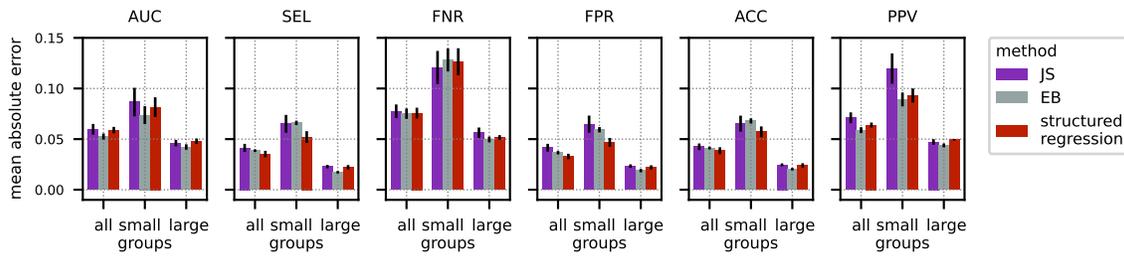
Figure 7: *Mean absolute error of estimates of 6 metrics using 3 best-performing methods on diabetes data.* **Averaged across all groups, small groups (size at most 25), and large groups (size above 25), across 20 draws of evaluation dataset.**
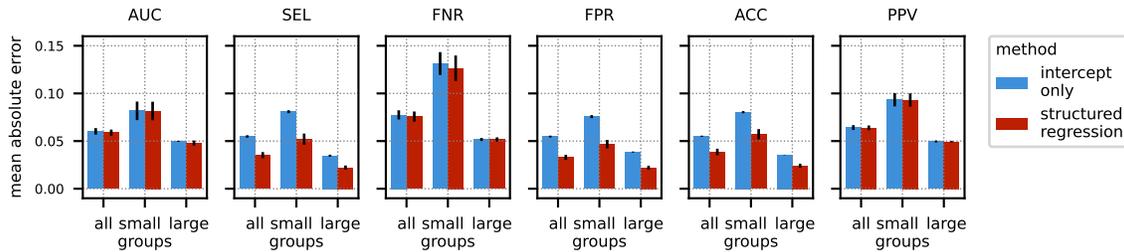
Figure 8: *Comparison of structured regression with the intercept-only model.* **Showing mean absolute error, averaged across all groups, small groups (size at most 25), and large groups (size above 25), across 20 draws of evaluation dataset.**

interaction terms. The former correspond to the situation when harms experienced by intersectional groups combine additively, the latter when there is an additional intersectional effect. For the additive ground truth (model$_{age+rc}$), tests suggest a sequence of variable additions *expl* + *rc* + *age*, but then show no support for including interaction terms. For the data from model$_{age\cdot rc}$, tests correctly provide support for an inclusion of interactions. That is,

we correctly identify the presence of super-additive harms that would accrue to intersectional age-race subgroups.

D ADDITIONAL DIABETES EXPERIMENTS

In Figure 7, we evaluate the quality of the point estimates produced by the three best-performing methods. In Figure 8, we compare the performance of the structured regression approach with an intercept-only model.