# Finding Regions of Heterogeneity in Decision-Making via Expected Conditional Covariance

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Individuals often make different decisions when faced with the same context, due to personal preferences and background. For instance, judges may vary in their leniency towards certain drug-related offenses, and doctors may vary in their preference for how to start treatment for certain types of patients. With these examples in mind, we present an algorithm for identifying types of contexts (e.g., types of cases or patients) with high inter-decision-maker disagreement. We formalize this as a causal inference problem, seeking a region where the assignment of decision-maker has a large causal effect on the decision. We give an iterative algorithm to find a region maximizing this objective and give a generalization bound for its performance. In a semi-synthetic experiment, we show that our algorithm recovers the correct region of disagreement accurately compared to baselines. Finally, we apply our algorithm to a real-world healthcare dataset, recovering variation that aligns with existing clinical knowledge.

## 1 Introduction

Understanding heterogeneity in decision-making is an established problem in medicine (Birkmeyer et al., 2013; Corallo et al., 2014; De Jong et al., 2006), consumer choice (Ortega et al., 2011; Scarpa et al., 2005), and law (Kang et al., 2012; Kleinberg et al., 2018; Arnold et al., 2018). In the context of medicine, this is referred to as the study of practice variation (Atsma et al., 2020), where it is often observed that doctors, facing the same clinical context, will make different decisions. Likewise, in a legal context, it is well documented that judges differ in their leniency in their decisions regarding bail (Kleinberg et al., 2018), juvenile incarceration (Aizer and Doyle Jr, 2013), the use of alternatives to incarceration (Di Tella and Schargrodsky, 2013), and incarceration length (Kling, 2006). In some scenarios this variation may be justified: The best medical treatment may not be obvious. In others, it may be grossly unfair, as in the case of racial bias in bail decisions (Arnold et al., 2018).

In this work, we tackle the question of how to find and characterize this variation in the first place. In particular, we present a learning algorithm for identifying types of contexts (e.g., patients, cases) for which the identity of the decision-maker makes a substantial difference in the ultimate decision. In a medical context, a better understanding of treatment variation can inform the development and dissemination of clinical guidelines. In the legal domain, characterizing the cases where judges vary most in their leniency may be of independent interest to investigate potential issues of fairness. Our algorithm does not directly suggest guidelines for how to reduce variation in the regions it discovers. In both of these domains, guidelines and structural changes are tightly regulated, and we anticipate much more careful study with domain experts before any societal impact is made.

We formalize characterizing the region of variation as a causal inference problem: We want to characterize examples where *changing the decision-maker* would have resulted in a different decision. The challenge is two-fold: First, we only observe a single decision-maker per example, so we cannot

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

know e.g., how multiple judges would have decided the same case. Second, our data on individual decision-makers is often scarce. For instance, in Section 5, we consider a medical dataset with more than 400 doctors, each of whom has fewer than 9 samples on average.

We will refer to decision-makers as "agents", and our contribution are as follows: In Section 2, we propose an objective defined in terms of counterfactual decisions across different agents, and show that this objective can be identified from observational data. Moreover, this objective does not require the use of agent-specific statistical models, making it amenable to our sparse setting. In Section 3, we give an iterative algorithm to identify regions of disagreement by maximizing this objective and provide intuition (in the form of a generalization bound) for the factors that drive its performance. In Section 4, we use a semi-synthetic dataset, derived from crowd-sourced recidivism predictions, to demonstrate that our algorithm recovers the correct region of disagreement accurately, even when there are many agents. Finally, in Section 5, we apply our algorithm to a real-world healthcare dataset and confirm that it recovers intuitive regions of variation in first-line diabetes treatment. We conclude with a discussion of related work and implications.

# 2 Characterizing Heterogeneity from a Causal Perspective

## 2.1 Notation

We consider our data to be drawn from a distribution $\mathbb{P}(X, A, Y)$, where $X$ is a random variable that represents context (or features), $A$ is a discrete variable that represents an agent, and $Y \in \{0, 1\}$ a binary variable that represents a decision. We denote the space of all $X$ as $\mathcal{X}$ and of all $A$ as $\mathcal{A}$. We use indicator variables, denoted as $\mathbf{1}[\cdot]$, which are equal to one if the statement inside the brackets is true and zero otherwise. For a subset such as $S \subseteq \mathcal{X}$, we will sometimes write the indicator $\mathbf{1}[x \in S]$ as a function $S(x)$, where $S : \mathcal{X} \to \{0, 1\}$. We use lower case letters (e.g., $x, a$) to denote realized values of random variables.

## 2.2 Heterogeneity as a Causal Contrast

Our conceptual goal is to identify a region $S \subseteq \mathcal{X}$ where different agents tend to make different decisions even when faced with the same context. We can formalize this in the language of potential outcomes from the causal inference literature (Pearl, 2009; Hernán and Robins, 2020), which for clarity we will refer to as *potential decisions*: In particular, we denote $Y(a)$ to be the potential decision of agent $a$. The fundamental challenge of causal inference is that we do not observe all potential decisions $\{Y(a) : a \in \mathcal{A}\}$ for each sample, but only a single decision $Y$. With this in mind, we will make the following assumptions, standard in the literature on causal effect estimation.

**Assumption 1** (Causal Identification Assumptions)**.** *(i) Consistency: $Y = y, A = a \implies Y(a) = y$, and (ii) No Unmeasured Confounding (NUC): For all $a \in \mathcal{A}$, $Y(a) \perp\!\!\!\perp A \mid X$.*

Consistency links the potential $Y(a)$ to the observed $Y$, and NUC says that there are no unobserved factors that influence both the assignment of agents and the decision itself. Judges, for instance, are often quasi-randomly assigned to cases (Kleinberg et al., 2018). Given these assumptions, we propose a causal measure of agent-specific bias, defined as a contrast between potential decisions.

**Definition 1** (Conditional Relative Agent Bias)**.** For an agent $a \in \mathcal{A}$, and a subset $S \subseteq \mathcal{X}$, the conditional relative agent bias is defined as

$$\mathbb{E}[Y(a) - Y(\pi(x)) \mid A = a, X \in S] \tag{1}$$

where $Y(a)$ is the potential decision of agent $a$, and $Y(\pi(x)) := \sum_{a'} \mathbb{E}[Y(a') \mid x]\pi(a' \mid x)$ denotes the expected potential decision under the agent assignment distribution $\pi(a' \mid x) := \mathbb{P}(A = a' \mid X = x)$.

We note that under Assumption 1, $Y(\pi(x)) = \mathbb{E}[Y \mid x]$,[1] but we write it this way to emphasize its causal interpretation as the expected decision of a random agent. Equation (1) represents the relative difference between the decision of an agent (on their particular distribution of cases) and the potential decision of a random agent. In particular, Equation (1) can be written as follows under Assumption 1

$$\mathbb{E}[Y(a) - Y(\pi) \mid A = a, X \in S] = \int_{x \in S} \mathbb{E}[Y(a) - Y(\pi) \mid X = x] p(x \mid a) dx, \tag{2}$$

[1] See Appendix A.1 for a short proof of this fact and for a derivation of Equation (2).

where we shorten $Y(\pi(x))$ to $Y(\pi)$. This is the average difference (over $p(x \mid a)$) of the conditional expected difference between $Y(a)$ and $Y(\pi)$. As a concrete example, suppose that the agent $a$ is a judge who is particularly lenient on bail decisions for felony arrests (the region $S$), and $Y = 1$ denotes granting bail. We then imagine the following counterfactual: We take the felony cases that are assigned to this judge, and we reassign each individual case, described by $x$, to a random judge $a'$, proportionally to $p(a' \mid x)$. We may then observe, on average, that the bail rate would decrease, because most judges are less lenient than judge $a$, corresponding to a positive value of Equation (1).

### 2.3 Formalizing a Causal Objective

Given our proposed measure of agent-specific bias on a region $S$, we now define an objective that captures aggregate bias across a grouping of agents. In particular, let the binary function $G : \mathcal{A} \to \{0, 1\}$ indicate a grouping of agents, for which we can define the following objective as both a function of the region $S$ and group $G$

$$Q(S, G) \coloneqq \sum_{a:G(a)=1} \mathbb{P}(A = a \mid x \in S)\mathbb{E}[Y(a) - Y(\pi) \mid A = a, X \in S] \tag{3}$$

This objective is maximized (for a given $S$) by choosing $G(a) = 1$ if the agent-specific bias is non-negative on $S$. We now show that this quantity can be identified and estimated from observational data without requiring agent-specific statistical models.

**Theorem 1** (Causal Identification)**.** *Under Assumption 1, $Q(S, G)$ can be identified as*

$$Q(S, G) = \mathbb{E}_S[Cov(Y, G \mid X)] = \mathbb{E}_S[(Y - \mathbb{E}[Y \mid X])G], \tag{4}$$

*where $\mathbb{E}_S[\cdot] \coloneqq \mathbb{E}[\cdot \mid X \in S]$, and $Cov(Y, G \mid X)$ is the conditional covariance.*

Theorem 1 and other theoretical results are proven in Appendix A. The result follows from proving that the agent-specific bias (Definition 1) is identifiable using the expected conditional covariance between $Y$ and the binary indicator $\mathbf{1}[A = a]$.

With this in mind, we optimize the following objective, where we constrain the set $S$ to be of at least a certain size $\beta$, and where $\mathcal{S}$ is our hypothesis class of functions $S$.

$$\max_{S \in \mathcal{S}, G} Q(S, G) \text{ s.t., } \mathbb{P}(S) \geq \beta, \tag{5}$$

**Remark**: We have *not* made the overlap or positivity assumption that $\mathbb{P}(A = a \mid x) > 0$ for all $x, a$. While this assumption is required to identify conditional causal effects of the form $\mathbb{E}[Y(a) - Y(a') \mid X]$ (Nie and Wager, 2017; Wager and Athey, 2018; Shalit et al., 2017), it is not required for identifying our causal contrast, which has a different interpretation. For instance, suppose that $S$ contains both violent and non-violent crimes in our example above. If judge $a$ makes bail decisions primarily for violent crimes, but some other judges *never* see violent crimes, then we cannot hope to compare $Y(a)$ to the decisions of those alternative judges. Instead, we have chosen to compare $Y(a)$ to the decisions of *viable alternative agents*, weighted by their probability $p(a' \mid x)$ of being selected for such a case.

## 3 Identifying Regions with Heterogeneity

In Section 3.1, we introduce an iterative optimization algorithm (Algorithm 1) which optimizes objective (5) by alternatingly optimizing $S$ and $G$. In Sections 3.2 and 3.3, we discuss practical approaches for choosing the region size parameter $\beta$ on training data and validating the significance of the resulting region on held-out data. Finally, we build intuition for the factors that influence performance of this algorithm via a generalization bound in Section 3.4.

### 3.1 Iterative Optimization Algorithm

We let $\hat{Q}(S, G)$ be the empirical analog of $Q(S, G)$ (Equation 4), which we can write as follows,

$$\hat{Q}(S, G) \coloneqq \frac{1}{\sum_{a,j} \mathbf{1}\left[x_{aj} \in S\right]} \sum_{a,j} (y_{aj} - f(x_{aj})) \cdot G(a) \cdot \mathbf{1}\left[x_{aj} \in S\right]. \tag{6}$$

**Algorithm 1** Identifying regions with variation

**Input:** Data $\{\{x_{aj}, y_{aj}\}_{j=1}^{R}\}_{a=1}^{N}$, minimum region size $\beta$.
Fit a model $f(x)$ to $\mathbb{E}(Y \mid X = x)$.
Initialize $S = \mathcal{X}$.
**repeat**
  **for** $a = 1, \ldots, N$ **do**
    Compute $\hat{Q}(S, \mathbf{1}[A = a])$, where $\hat{Q}(S, \mathbf{1}[A = a]) := \frac{1}{\sum_j \mathbf{1}[x_{aj} \in S]} \sum_{x_{aj} \in S} (y_{aj} - f(x_{aj}))$,
    Set $G(a) = 1$ if $\hat{Q}(S, \mathbf{1}[A = a]) \geq 0$ and 0 otherwise.
  **end for**
  Compute $b_{aj} = (y_{aj} - f(x_{aj}))G(a)$, $a = 1, \ldots, N$, $j = 1, \ldots, R$.
  Fit a model $h(x)$ to predict $b_{aj}$ from $x_{aj}$, and let $b$ be the $(1 - \beta)$-th quantile of $h(x_{aj})$.
  $S' \leftarrow S$.
  $S \leftarrow \{x_{aj};\ h(x_{aj}) \geq b\}$.
**until** $S = S'$ or iteration limit reached.
**Output:** Model $h$ and threshold $b$, defining a region $S = \{x \in \mathcal{X}; h(x) \geq b\}$.

where $f(x)$ denotes a model of the conditional expectation $f(x) \approx \mathbb{E}[Y \mid X = x]$. For simplicity of notation, we assume that there are $R$ samples (indexed by $j$) for each of a finite set of $N$ agents (indexed by $a$), giving $N \cdot R$ samples in total.

Our algorithm takes as input the data $\{(x_{aj}, y_{aj})\}$ and a minimum region size $\beta$, and outputs a model $h(x)$ and a threshold value $b$, describing a region of disagreement $S = \{x \in \mathcal{X}; h(x) \geq b\}$. Starting with $S = \mathcal{X}$ (the entire space), the algorithm identifies the grouping that maximizes $\hat{Q}(S, G)$, then uses that grouping to identify the region maximizing the same quantity. This process is then repeated until convergence. The algorithm uses a classifier $f(x)$ to estimate the quantity $\mathbb{E}[Y \mid X = x]$ and a regression model $h(x)$ to estimate the conditional covariance of the decision $Y$ and the grouping $G$ at $X = x$. Note that we can use any supervised learning algorithms for $f$ and $h$, allowing us to learn interpretable regions as part of the algorithm if $h(x)$ is interpretable (e.g., decision trees).

**Optimizing over $G$ given $S$.** Given a region $S$, our first result identifies the grouping $G : \mathcal{A} \to \{0, 1\}$ that maximizes $\hat{Q}(S, G)$ and shows that it can be expressed in terms of $\hat{Q}(S, \mathbf{1}[A = a])$.

**Proposition 1.** *Given $S \subseteq \mathcal{X}$, $\hat{Q}(S, G)$ is maximized over the space of functions $G : \mathcal{A} \to \{0, 1\}$ at $G_S$, where $G_S(a) = \mathbf{1}\left[\hat{Q}(S, \mathbf{1}[A = a]) \geq 0\right]$.*

Intuitively, this proposition states that to maximize the empirical expected conditional covariance of the decision and grouping on a region, we must group agents by whether their residuals $y_{aj} - f(x_{aj})$ are (on average) positive or negative on $S$.

**Optimizing over $S$ given $G$.** To optimize $\hat{Q}(S, G)$ for a fixed grouping $G$ over the hypothesis class $\mathcal{S}$, we train a model $h(x)$ to predict $(y_{aj} - f(x_{aj}))G(a)$ given $x_{aj}$, where $h \in \mathcal{H}$. Using $h$ as an estimate in Eq. 6, we now find a set $S$ to maximize the quantity $\frac{1}{\sum_x \mathbf{1}[x \in S]} \sum_{x \in S} h(x)$, the empirical conditional expectation of $h(x)$ over $S$. This quantity is maximized (subject to our $\beta$ constraint) by including the largest $\beta$-fraction of the $h(x_{aj})$ in $S$. Hence, we pick $b$ as the $(1 - \beta)$-th quantile of $h(x_{aj})$ and choose our region as $\hat{S}_G = \{x \in \mathcal{X}; h(x) \geq b\}$.

### 3.2 Tuning the Region Size Parameter

For real datasets, we need to choose $\beta$ without knowledge of the true value. Given that our objective can be calculated on held-out data using the functions $S, G$, a seemingly obvious approach would be to compute $Q(S, G)$ on a validation set and select the $\beta$ that leads to the highest $Q(S, G)$. However, for a fixed data distribution, smaller values of $\beta$ tend to produce higher values of $Q(S, G)$, and there is a trade-off between finding a smaller region of higher variation and a larger region that may include areas of lower (but still meaningful) variation. This motivated our original constraint $\mathbb{P}(S) \geq \beta$.

To facilitate selection of $\beta$, we propose a heuristic inspired by permutation-based hypothesis testing (Wasserman, 2004). The conceptual idea is to compare the training objective (for a fixed $\beta$) to a reference distribution of values (for the same $\beta$) that might be seen *even if all agents followed the same policy*. For each candidate $\beta$, we (i) run our algorithm and compute the objective value on training

data $q_{\text{obs}} := \hat{Q}(\hat{S}, \hat{G})$. (ii) For $T$ iterations, we randomly shuffle the agents and re-run the algorithm to get a new objective value. This gives us a distribution $\hat{\mathbb{P}}_{\text{null}}$ over $Q(S, G)$ from a distribution where $\mathbb{P}(X, Y)$ and $\mathbb{P}(A)$ are unchanged but $Y, X \perp\!\!\!\perp A$. (iii) Finally, we compute a p-value $p_\beta = \hat{\mathbb{P}}_{\text{null}}(Q > q_{\text{obs}})$ and choose the $\beta$ with the smallest p-value. In Section 4.2, we find that this heuristic empirically recovers the true $\beta$ value in semi-synthetic settings.

### 3.3 Validation of the Region

In real-data experiments, we may wish to validate the learned region $\hat{S}$ independently of the grouping $\hat{G}$. In particular, $\hat{G}$ may not be of intrinsic interest, and we observe in our semi-synthetic experiments that our algorithm can achieve high AUC in finding the true region $S$, even when the accuracy of the grouping $G$ is fairly poor (due to small numbers of samples per agent), as shown in Appendix B.2. Thus, we optimize out the $G$ from the objective $Q(S, G)$ to obtain an objective that depends only on $S$ and can be used to compare different regions directly. Using Proposition 1, we write

$$\hat{L}(S) := \max_G \hat{Q}(S, G) = \frac{1}{\sum_{a,j} \mathbf{1}\left[x_{aj} \in S\right]} \sum_a \left[ \sum_j (y_{aj} - f(x_{aj})) \mathbf{1}\left[x_{aj} \in S\right] \right]_+, \tag{7}$$

where $[\cdot]_+$ is equal to the term inside the brackets if that term is positive and zero otherwise. We then use this objective $\hat{L}(S)$ to answer the following question: Does our chosen region $\hat{S}$ yield a significantly higher objective value on test data than a randomly selected region of the same size? An example of this analysis is given in Table 1 for the real-data experiment in Section 5.

### 3.4 Generalization Error

In this section, we give a generalization bound for Algorithm 1, showing that it can recover, with high probability, a region whose value $Q(S, G^*)$ is close to that of the optimal $S^*$. Recall that $S^*$ and $G^*$ are defined as the maximizers of Objective (5). We provide this bound to help develop intuition, as it depends upon a few simplifying assumptions:

**Assumption 2** (Group-based variation)**.** *For all $x \in S^*$, $\mathbb{E}[Y \mid X = x, A = a] = \mathbb{E}[Y \mid X = x, G^*(a)]$ and for all $x \notin S^*$, $\mathbb{E}[Y \mid X = x, A = a] = \mathbb{E}[Y \mid X = x]$*

**Assumption 3** (Non-zero relative biases)**.** *There exists a constant $\alpha > 0$ such that for all $x \in S^*$,*

$$\mathbb{E}[Y \mid X = x, G^*(A) = 1] - \mathbb{E}[Y \mid X = x] > \alpha, \quad \textit{and} \quad \mathbb{E}[Y \mid X = x, G^*(A) = 0] - \mathbb{E}[Y \mid X = x] < -\alpha,$$

**Assumption 4** (Coverage)**.** *There exists a constant $\omega > 0$, such that for every $a \in \mathcal{A}$, and for all $x \in S^*$, $\mathbb{P}(X = x \mid A = a)/\mathbb{P}(X = x) > \omega$.*

Assumption 2 says that there are two groups of agents, who follow two distinct decision policies within a region $S^*$, but follow an identical decision policy outside of $S^*$. Assumption 3 says that one group has a positive bias across all of $S^*$, relative to the average over both groups, and the other group has a negative bias.To simplify the analysis, we also make Assumption 4 that every agent has some non-zero chance of observing every value of $x$ in $S^*$. These assumptions are not required for the identification of our objective (see Theorem 1). For simplicity, we assume that $f(x)$ perfectly recovers $\mathbb{E}[Y \mid X]$. This can be relaxed at the cost of additional terms in the bound that go to zero as the overall sample size increases. Under Assumptions 2, 3, and 4, assume that $\mathbb{P}(S^*) = \beta$. For the informal version presented here, we assume that $\mathbb{P}(S^*) = \mathbb{P}(\hat{S}) = \beta$, where $\hat{S}$ is returned by our algorithm, and that exactly a $\beta$-fraction of our samples fall into $S^*$ and $\hat{S}$ (in the Appendix, we give a version without these simplifications).

**Theorem 2.** *Under the assumptions above, if $R > \frac{2 \ln 2}{\alpha^2 \beta^2 \omega^2}$, the first iteration of Algorithm 1 returns $\hat{S}$ such that, with probability at least $1 - \delta$, $Q(S^*, G^*) - Q(\hat{S}, G^*) \leq \epsilon$, where*

$$\epsilon = \sqrt{\frac{2 \ln(3/\delta)}{\beta N \cdot R}} + \frac{2}{\beta}\left(\eta + \sqrt{\frac{3\eta(1-\eta)}{\delta \cdot N}}\right) + \frac{1}{\beta}\left(2\mathcal{R}(\mathcal{S}, N \cdot R) + 4\sqrt{\frac{2 \ln(12/\delta)}{N \cdot R}}\right),$$

*where $\mathcal{R}(\mathcal{S}, N \cdot R)$ is the Rademacher complexity of $\mathcal{S}$, and*

$$\eta = \exp\left(\frac{-R\alpha^2\beta^2\omega^2}{2}\right).$$

The term $\eta$ plays an important role: It bounds the expected misclassification error $\mathbb{P}(\hat{G}(a) \neq G^*(a))$. For sufficiently large $R$, we have that $\eta < 1/2$ with high probability, i.e., we have a better-than-random chance of identifying the group for an individual agent. Moreover, $\eta$ decreases as we increase the number of samples $R$ for each agent, the separation $\alpha$ between the two groups on $S^*$, the region size $\beta$, and the coverage $\omega$ of agents over $S^*$. For sufficiently small $\eta$, our algorithm discovers a region whose value (in terms of $Q(S, G^*)$) is close to that of the optimal region. The generalization bound improves as the number of agents $N$ increases, the number of samples $R$ for each agent increases, or the complexity of the hypothesis class decreases. The latter is measured here by the Rademacher complexity $\mathcal{R}(\mathcal{S}, N \cdot R)$ of our hypothesis class $\mathcal{S}$, which can be bounded by standard arguments. In conclusion, under some additional assumptions, Algorithm 1 identifies an approximately optimal solution with high probability after one iteration. We show via semi-synthetic experiments in Appendix B.3 that convergence is generally fast in practice.

# 4 Semi-Synthetic Experiment: Recidivism Prediction

For conceptual motivation in the introduction, we discussed the legal system: As a potential application of our method, one could determine types of cases for which the idiosyncratic preferences of judges have a significant impact on their decisions. Lacking data on judge decisions with sufficient context, we turn to the more controlled setting of human predictions of recidivism.

**Dataset:** We use publicly available data from Lin et al. (2020),[2] who ask participants on Amazon's Mechanical Turk platform to make recidivism predictions based on information present in the "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) dataset for Broward County, FL (Dressel and Farid, 2018). Participants (or "agents") are provided with 5 risk factors to make their predictions: age, gender, and number of prior convictions, juvenile felony charges, and juvenile misdemeanor charges. The charge in question is also given, as well as whether the charge is a misdemeanor or felony. The dataset contains 4550 cases evaluated by 87 participants.

**Semi-Synthetic Policy Generation:** To benchmark our method, we generate semi-synthetic data where we have access to a "ground truth" region of variation. We retain the features presented to the original participants, and construct two policies, which we refer to as the "base" and "alternative" policy: For the first policy, we learn a logistic regression model on the binary decisions across the whole dataset. For the alternative policy, an extra positive term is added to the logistic regression for samples within the region. We construct two scenarios with different regions of variation: (1) all drug possession charges, and (2) all misdemeanor charges where the individual is 35 years old or younger. These make up 22% and 21% of the data, respectively. Then, we generate synthetic agents (randomly assigned to cases) and assign half of the agents to the base policy and half to the alternative. Synthetic decisions are then sampled from the logistic regressions. For each scenario, the two groups of agents follow the same stochastic policy outside of the region, and one group systematically prefers $Y = 1$ within the region. More details can be found in Appendix B.1.

## 4.1 Performance versus Baselines

**Baselines:** To evaluate our approach, we compare to a series of baselines approaches, on the basis of performance in identifying the true region of variation. To our knowledge, the problem of finding regions of heterogeneity with a large number of agents has not been studied before, so we develop new baselines. Some (including those described in Appendix B.2) are based on causal inference methods augmented with a way to identify a region of variation and grouping of agents where possible.

*Direct models:* This baseline measures how much adding the agent to the feature set improves prediction of decisions. We fit logistic regressions with and without the agent feature to estimate $\mathbb{E}[Y \mid A, X]$ and $\mathbb{E}[Y \mid X]$. For each sample $(x, y, a)$, we compute $|y - \mathbb{E}[Y \mid X = x]| - |y - \mathbb{E}[Y \mid A = a, X = x]|$ to quantify how much the model with agents outperforms the model without agents. Then, we fit a "region model" to predict this quantity from $X$. This region model is either a ridge regression, decision tree, or random forest model. Finally, we compute the top $\beta$ quantile of predictions from the region model in the training and validation sets and use this cut-off to select points in the test set.

*TARNet:* A treatment-agnostic representation network (Shalit et al., 2017) models the outcomes of all treatments for each sample by learning a shared representation of the sample features and then having

[2] Available at `https://github.com/stanford-policylab/recidivism-predictions`

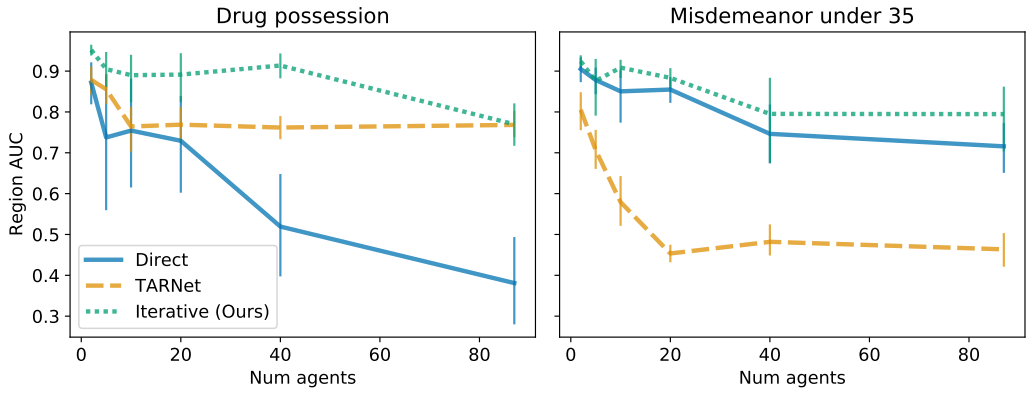



Figure 1: Comparison of our method and best baselines at identifying region of disagreement, as measured by AUC of the region model on a held-out test set. Total number of samples is fixed. Direct refers to the baseline that models $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[Y \mid X, A]$. Uncertainty bands represent 95% intervals for the mean derived via bootstrapping over 10 random seeds using seaborn (Waskom, 2021). Left: Region is modelled using a ridge regression in the drug possession semi-synthetic set-up. Right: Region is modelled using a random forest for the misdemeanor under age 35 set-up.

a separate prediction head for each treatment. We implement the shared representation model using a 2-layer neural network with ReLU and dropout. Each prediction head is a linear layer, and we use the binary cross entropy loss function from PyTorch (Paszke et al., 2019) that includes a sigmoid function. TARNet predicts $\mathbb{E}[Y \mid X, A]$ for every agent for each sample. We compute $\mathrm{Var}_A[\mathbb{E}[Y \mid X, A]]$ and then predict this quantity with the region models as in the direct model baseline. Additional baseline details are in Appendix B.2.

**Results:** We evaluate how well the algorithms identify the samples within the region of variation when we vary the number of agents among 2, 5, 10, 20, 40, and 87, where 87 is the number of agents in the original dataset. Figure 1 shows the best overall region models for each set-up, with the other models deferred to Appendix B.2. Algorithm 1 consistently performs well for both semi-synthetic set-ups, especially when the number of agents increases to a realistic level (and the number of samples per agent decreases). The direct baseline deteriorates very rapidly as the number of agents increases in the drug possession set-up, while the TARNet baseline deteriorates rapidly in the misdemeanor under age 35 set-up. Refer to Appendix B.2 for additional baselines, region models, and evaluation metrics. We also show that our method is robust in a set-up with more than 2 agent groups in Appendix B.4.

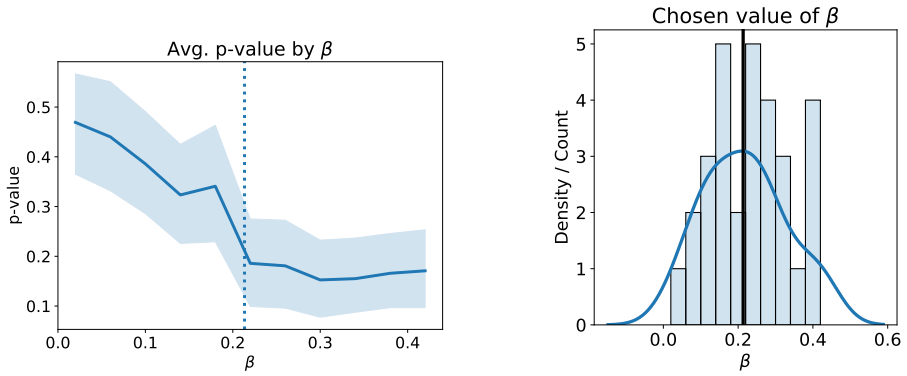
### 4.2 Tuning the Region Size

We validate the proposed approach of tuning $\beta$ (discussed in Section 3.2), by applying the methodology to our semi-synthetic setting here. We sample 30 semi-synthetic datasets and consider candidate values of $\beta$ in $[0.02, 0.42]$ in increments of 0.04. For each proposed value of $\beta$ we use $T = 40$ random permutations of the agents. Figure 2 presents results for the misdemeanor under age 35 set-up. As the candidate value of $\beta$ increases (up to the true value of $\beta$), the p-value decreases, and the distribution of selected $\beta$ values are centered on the true value of $\beta$.

## 5 Real-data Experiment: First-Line Diabetes Treatment

We apply our algorithm on a real-world dataset consisting of first-line (initial) treatment for type 2 diabetes and examine how the treatment variation we discover aligns with clinical knowledge. We present an additional real-world experiment (on Parkinson's disease) in Appendix D.

**Data and Setup**: We use an observational dataset consisting of medical claims from a large insurance provider, spanning from 2010 to 2020. The task is to classify first-line treatment decisions between metformin ($Y = 0$) and other common first-line treatments such as sitagliptin, glipizide, glimepiride, or glyburide ($Y = 1$) (Hripcsak et al., 2016). As relevant clinical features, we include the patients'




(a) (b)

Figure 2: Results of tuning $\beta$ over 30 semi-synthetic datasets. (a) Average p-value for each candidate $\beta$, with 95% uncertainty intervals for the mean generated by bootstrapping. The dotted vertical line represents the true value of $\beta$. (b) Distribution of $\beta$ with the smallest p-value over the datasets.

Table 1: Objective values $L(S)$ for the learned region on the training and test datasets, along with the distribution of values for randomly generated regions $S_{\text{rand}}$.

| **Metric** | **Subset** | **Value** |
|---|---|---|
| $L(\hat{S})$ | Train | 0.1029 |
| $L(\hat{S})$ | Test | 0.0924 |
| $L(S_{\text{rand}})$ | Test | 0.0507 (0.0073) |

most recent eGFR (mL/min/1.73m2) and creatinine (mg/dL) measurements, incidence of heart failure, and treatment date. Metformin is typically recommended by guidelines, unless the patient has poor kidney function, typically indicated by low eGFR or high creatinine (American Diabetes Association, 2010). The agent indicator $A$ is the group practice of the doctor responsible for the patient's treatment. 3,956 patients and 458 group practices are included in our cohort. In this experiment, we choose $\beta = 0.25$ as input to our algorithm.

**Interpretation of Results**: To interpret the region, we use decision trees as our region model $h(x)$. Other model details, as well as a visualization of the resulting tree, are available in Appendix C. The decision tree identifies the region of disagreement as the union of (i) eGFR below 58.5, (ii) treatment date before February 2014, with eGFR above 95.5 and creatinine below 1.11, and (iii) treatment date after February 2014, with eGFR above 58.5 and creatinine between 1.01 and 1.11. These regions align with clinical intuition. In the first region, low eGFR values indicate impaired renal function (Group et al., 2009), which is a contraindication for metformin since it is traditionally associated with increased risk of lactic acidosis (Tahrani et al., 2007). Still, treatment decisions can vary here because guidelines for managing patients with eGFR below 45 are lacking (Group et al., 2015); for instance, one study found that metformin use was more prevalent than was recommended for patients with eGFR below 60 (Vasisht et al., 2010). The second and third regions split by treatment date, and include patients who are not obviously contraindicated for metformin. Variation here may be associated with the increasing use of metformin among doctors over time (Landon et al., 2018).

**Assessing Significance**: In Table 1, we perform a sanity check, assessing whether our algorithm identifies a region $S$ whose variation in held-out data is higher than that of a randomly selected region, using the partially optimized objective $L(S) = \max_G Q(S, G)$ laid out in Section 3.3 to compare regions directly. In Table 1, we compare $L(\hat{S})$ on training and test data to the distribution of $L(S_{\text{rand}})$, where $S_{\text{rand}}$ are random subsets of data of the same size as $\hat{S}$. We compute the latter distribution using 100 random subsets, and observe that the test value of $L(\hat{S})$ is more than two standard deviations above the mean of the latter. This gives us confidence that the discovered region $S$ generalizes to a region of disagreement beyond the training set. We direct the reader to Appendix C for additional analyses, such as evaluating the stability of the region over multiple folds of splitting the data.

## 6 Related Work

Beyond previously mentioned connections to causal effect estimation, we highlight a few areas of research that share similar goals to our own.

**Conditional independence testing**: While we seek to find a region that maximizes a causal notion dependence, one could instead ask whether or not $Y$ is conditionally independent of $A$ given $X$. A variety of metrics exist for testing conditional independence, such as the Hilbert-Schmidt independence criterion (HSIC) (Fukumizu et al., 2007; Zhang et al., 2012), conditional mutual information (Runge, 2018), and conditional correlation (Ramsey, 2014). We make use of the expected conditional covariance, which has also been proposed for conditional independence testing (Shah and Peters, 2020), showing that it can be given a causal interpretation under certain assumptions. We seek a region that maximizes this measure, in lieu of using it for testing.

**Hierarchical / Mixture models**: Bayesian methods are often used to estimate group-level effects, such as a per-physician effect on patient outcomes (Tuerk et al., 2008), where group identifiers are included as a categorical feature in a multi-level generalized linear model (Gelman and Hill, 2006). Alternatively, one could assume a conditional mixture model (Bishop, 2006), where agents belong to latent clusters that each have their own policy. However, both of these methods require parametric assumptions on the distribution of $\mathbb{P}[Y \mid x, a]$, and even so, the optimal mixture model is not necessarily identifiable when both the clusters and policies are unknown (Dasgupta and Schulman, 2007). By contrast, our method does not require making particular parametric assumptions about $\mathbb{E}[Y \mid x, a]$ and seeks to learn the region of variation directly.

**Feature evaluation**: Checking for heterogeneity can also be framed as feature evaluation, where we would like to evaluate whether adding the agent identifier will increase predictive power. Feature evaluation methods typically maximize dependence between selected features and labels, utilizing measures similar to those in conditional independence testing, such as the HSIC (Song et al., 2007). Other methods use the correlation of the new feature with the loss gradient of the current predictor as a measure of utility (Koepke and Bilenko, 2012). In contrast, we focus not on checking marginally for the predictive power of agent identifiers, but rather identifying a region.

**Crowdsourcing**: Our work is conceptually related to identifying which samples are difficult to label via crowdsourcing annotations (Karger et al., 2014; Whitehill et al., 2009). Most crowdsourcing models are generative with latent variables for the correct sample labels, sample difficulty, agent expertise, and agent bias. They then optimize for the likelihood of the observed labels. The set of difficult samples is analogous to our region of disagreement. The main difference with our problem is that we do not require any notion of the "correct" label.

## 7 Discussion

In this work, we take a causal perspective on finding regions where agents (i.e., decision-makers) have heterogeneous preferences, formalizing this heterogeneity in terms of counterfactual (or "potential") decisions across agents. We propose a causal measure of agent-specific biases and give an objective that aggregates this bias over agents. This objective can be identified from observational data and written in terms of an expected conditional covariance. Importantly, for our applications, this does not require building agent-specific models or assuming overlap across all agents.

We give an iterative algorithm to find a region that maximizes this objective. Then, we demonstrate in semi-synthetic experiments that our algorithm accurately recovers the region of variation and scales well as the number of agents increases. This is in contrast to baseline approaches we consider, where performance deteriorates when the number of agents increases. Finally, on a real-world medical dataset, we show that our algorithm can yield insights that align with clinical knowledge.

Our work is motivated by understanding variation in human decision-making. In the judicial domain, our method may help unearth types of cases where decisions are highly dependent on the judge assigned to the case. In the medical domain, our approach may identify types of patients where new guidelines may be required to help doctors make more uniform decisions. Our works seeks to discover variation: Domain expertise is required to determine the implications of the discovered region. Beyond these applications, we see our approach as a useful data science tool for understanding heterogeneity in decisions that appears to be driven by individual-level preferences.

## References


Anna Aizer and Joseph J Doyle Jr. 2013. Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges. Technical Report w19102. National Bureau of Economic Research.

American Diabetes Association. 2010. Standards of medical care in diabetes—2010. Diabetes care 33, Supplement 1 (2010), S11–S61.

David Arnold, Will Dobbie, and Crystal S Yang. 2018. Racial Bias in Bail Decisions. The Quarterly Journal of Economics 133, 4 (May 2018), 1885–1932.

Femke Atsma, Glyn Elwyn, and Gert Westert. 2020. Understanding unwarranted variation in clinical practice: a focus on network effects, reflective medicine and learning health systems. International Journal for Quality in Health Care 32, 4 (jun 2020), 271–274. https://doi.org/10.1093/intqhc/mzaa023

John D Birkmeyer, Bradley N Reames, Peter McCulloch, Andrew J Carr, W Bruce Campbell, and John E Wennberg. 2013. Understanding of regional variation in the use of surgery. The Lancet 382, 9898 (2013), 1121–1129.

Christopher M Bishop. 2006. Pattern recognition and machine learning. springer.

Ashley N Corallo, Ruth Croxford, David C Goodman, Elisabeth L Bryan, Divya Srivastava, and Therese A Stukel. 2014. A systematic review of medical practice variation in OECD countries. Health Policy 114, 1 (2014), 5–14.

Sanjoy Dasgupta and Leonard J Schulman. 2007. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. Journal of Machine Learning Research 8 (2007), 203–226.

Judith D De Jong, Gert P Westert, Ronald Lagoe, and Peter P Groenewegen. 2006. Variation in hospital length of stay: do physicians adapt their length of stay decisions to what is usual in the hospital where they work? Health Services Research 41, 2 (2006), 374–394.

Rafael Di Tella and Ernesto Schargrodsky. 2013. Criminal Recidivism after Prison and Electronic Monitoring. The Journal of Political Economy 121, 1 (2013), 28–73.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. Science advances 4, 1 (Jan. 2018), eaao5580.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2007. Kernel measures of conditional dependence.. In NIPS, Vol. 20. 489–496.

Andrew Gelman and Jennifer Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Christopher G Goetz, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Glenn T Stebbins, Matthew B Stern, Barbara C Tilley, Richard Dodel, Bruno Dubois, et al. 2007. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): process, format, and clinimetric testing plan. Movement disorders 22, 1 (2007), 41–47.

Guideline Development Group, Henk Bilo, Luis Coentrão, Cécile Couchoud, Adrian Covic, Johan De Sutter, Christiane Drechsler, Luigi Gnudi, David Goldsmith, James Heaf, et al. 2015. Clinical practice guideline on management of patients with diabetes and chronic kidney disease stage 3b or higher (eGFR< 45 mL/min). Nephrology Dialysis Transplantation 30, suppl_2 (2015), ii1–ii142.

Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBD Work Group et al. 2009. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). Kidney international. Supplement 113 (2009), S1–S130.

PD Med Collaborative Group et al. 2014. Long-term effectiveness of dopamine agonists and monoamine oxidase B inhibitors compared with levodopa as initial treatment for Parkinson's disease (PD MED): a large, open-label, pragmatic randomised trial. The Lancet 384, 9949 (2014), 1196–1205.

M A Hernán and J M Robins. 2020. Causal Inference: What If. Chapman & Hall/CRC, Boca Raton.

George Hripcsak, Patrick B Ryan, Jon D Duke, Nigam H Shah, Rae Woong Park, Vojtech Huser, Marc A Suchard, Martijn J Schuemie, Frank J DeFalco, Adler Perotte, et al. 2016. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences 113, 27 (2016), 7329–7336.

Jerry Kang, Judge Mark Bennett, Devon Carbado, Pam Casey, Nilanjana Dasgupta, David Faigman, Rachel Godsil, Anthony G. Greenwald, Justin Levinson, and Jennifer Mnookin. 2012. Implicit bias in the courtroom. UCLA Law Review 59, 5 (2012), 1124–1186.

David R Karger, Sewoong Oh, and Devavrat Shah. 2014. Budget-optimal task allocation for reliable crowdsourcing systems. Operations Research 62, 1 (2014), 1–24.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. The Quarterly Journal of Economics 133, 1 (Feb. 2018), 237–293.

Jeffrey R Kling. 2006. Incarceration Length, Employment, and Earnings. The American economic review 96, 3 (June 2006), 863–876.

Hoyt Koepke and Mikhail Bilenko. 2012. Fast prediction of new feature utility. arXiv preprint arXiv:1206.4680 (2012).

Bruce E Landon, Alan M Zaslavsky, Jeffrey Souza, and John Z Ayanian. 2018. Trends in diabetes treatment and monitoring among Medicare beneficiaries. Journal of general internal medicine 33, 4 (2018), 471–480.

Zhiyuan (Jerry) Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The Limits of Human Predictions of Recidivism. Science Advances 6, 7 (2020). https://doi.org/10.1126/sciadv.aaz0652 arXiv:https://advances.sciencemag.org/content/6/7/eaaz0652.full.pdf

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. 2011. The parkinson progression marker initiative (PPMI). Progress in neurobiology 95, 4 (2011), 629–635.

Sharon Muzerengi and Carl E Clarke. 2015. Initial drug treatment in Parkinson's disease. bmj 351 (2015).

Xinkun Nie and Stefan Wager. 2017. Quasi-oracle estimation of heterogeneous treatment effects. arXiv preprint arXiv:1712.04912 (2017).

David L Ortega, H Holly Wang, Laping Wu, and Nicole J Olynk. 2011. Modeling heterogeneity in consumer preferences for select food safety attributes in China. Food Policy 36, 2 (2011), 318–324. https://www.sciencedirect.com/science/article/pii/S0306919210001442

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019).

Judea Pearl. 2009. Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

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. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.

Joseph D Ramsey. 2014. A scalable conditional independence test for nonlinear, non-gaussian data. arXiv preprint arXiv:1401.5031 (2014).

Microsoft Research. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.x.

Jakob Runge. 2018. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In International Conference on Artificial Intelligence and Statistics. PMLR, 938–947.

Riccardo Scarpa, George Philippidis, and Fiorenza Spalatro. 2005. Product-country images and preference heterogeneity for Mediterranean food products: A discrete choice framework. Agribusiness 21, 3 (2005), 329–349.

Rajen D. Shah and Jonas Peters. 2020. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. arXiv:1804.07203 [math.ST]

Shai Shalev-Shwartz and Shai Ben-David. 2014. Understanding Machine Learning: From Theory to Algorithms. Vol. 9781107057. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning. PMLR, 3076–3085.

Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. 2007. Supervised feature selection via dependence estimation. In Proceedings of the 24th international conference on Machine learning. 823–830.

AA Tahrani, GI Varughese, JH Scarpello, and FWF Hanna. 2007. Metformin, heart failure, and lactic acidosis: is metformin absolutely contraindicated? Bmj 335, 7618 (2007), 508–512.

Peter W Tuerk, Martina Mueller, and Leonard E Egede. 2008. Estimating physician effects on glycemic control in the treatment of diabetes: methods, effects sizes, and implications for treatment policy. Diabetes care 31, 5 (May 2008), 869–873.

KP Vasisht, S-C Chen, Y Peng, and GL Bakris. 2010. Limitations of metformin use in patients with kidney disease: are they warranted? Diabetes, Obesity and Metabolism 12, 12 (2010), 1079–1083.

Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Amer. Statist. Assoc. 113, 523 (2018), 1228–1242.

Michael L. Waskom. 2021. seaborn: statistical data visualization. Journal of Open Source Software 6, 60 (2021), 3021. https://doi.org/10.21105/joss.03021

Larry Wasserman. 2004. All of Statistics: A Concise Course in Statistical Inference. Springer, New York, NY.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in neural information processing systems 22 (2009), 2035–2043.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2012. Kernel-based conditional independence test and application in causal discovery. arXiv preprint arXiv:1202.3775 (2012).

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We present the formalized problem and solution we describe in the abstract. We also run experiments that support the application examples we mention in the abstract.
   (b) Did you describe the limitations of your work? [Yes] Assumption 1 must hold.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] In Theorem 1, we collect the full set of assumptions in Assumption 1, and for Theorem 2, we give the relevant assumptions immediately before the statement of the theorem in Section 3.4.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs can be found in Section A of the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will submit code and instructions with the supplement. The recidivism and Parkinson's data are publicly available. The only exception is the diabetes data is proprietary from an insurance company collaborator.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.2 for how the region size hyperparameter is tuned, Appendix B.2 for how other hyperparameters are tuned, and other details in the paper.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See our figures.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We discuss this in the upfront overview section of the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite Lin et al. (2020) for the recidivism dataset. We cite Marek et al. (2011) for the Parkinson's dataset. The insurance data is proprietary.
   (b) Did you mention the license of the assets? [Yes] For the recidivism dataset, the data repository does not have a license at `https://github.com/stanford-policylab/recidivism-predictions`. For the Parkinson's dataset, we include the disclaimer from PPMI.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We do not curate/release any new assets, other than code for replicating our experiments; This code will be available in the supplementary material, as discussed above.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix C regarding consent for the proprietary health insurance claims dataset. All other data we use is publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix C regarding de-identification of the health insurance dataset. All other data we use is publicly available.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] The crowdsourced dataset was collected by another paper and publicly released.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]