Characterization of Overlap in Observational Studies

Michael Oberst* MIT-IBM Watson AI Lab, MIT, CSAIL & IMES
Fredrik D. Johansson* Chalmers University of Technology
Dennis Wei* MIT-IBM Watson AI Lab, IBM Research
Tian Gao MIT-IBM Watson AI Lab, IBM Research
Gabriel Brat Harvard Medical School
David Sontag MIT-IBM Watson AI Lab, MIT, CSAIL & IMES
Kush R. Varshney MIT-IBM Watson AI Lab, IBM Research

Abstract

Overlap between treatment groups is required for non-parametric estimation of causal effects. If a subgroup of subjects always receives the same intervention, we cannot estimate the effect of intervention changes on that subgroup without further assumptions. When overlap does not hold globally, characterizing local regions of overlap can inform the relevance of causal conclusions for new subjects, and can help guide additional data collection. To have impact, these descriptions must be interpretable for downstream users who are not machine learning experts, such as policy makers. We formalize overlap estimation as a problem of finding minimum volume sets subject to coverage constraints and reduce this problem to binary classification with Boolean rule classifiers. We then generalize this method to estimate overlap in off-policy policy evaluation. In several real-world applications, we demonstrate that these rules have comparable accuracy to black-box estimators and provide intuitive and informative explanations that can inform policy making.

1 INTRODUCTION

To accurately estimate the causal effect of an intervention, it is essential that intervention alternatives have been observed in comparable contexts, i.e., that there is *overlap* between the distributions of individuals receiving each intervention (Rosenbaum and Rubin, 1983; D'Amour et al., 2017). In randomized experiments, overlap is guaranteed for the study population by randomizing the intervention. However, this is not the case in observational studies where interventions are chosen according to an existing, in some cases deterministic, policy. In such settings, overlap may hold only for an unidentified subset of cases, with the causal effect being unidentifiable outside of this subset. We motivate our paper with the following use cases:

Scenario 1: From study to policy. When researchers publish the findings of a clinical trial, they also share the eligibility criteria (e.g., *Age $\geq$ 18, Serum M protein $\geq$ 1g/dl or Urine M protein $\geq$ 200 mg/24 hrs, Recent diagnosis* (National Cancer Institute, 2012)) and cohort statistics in order to characterize the cohort of study subjects. This gives policy makers means to assess the external validity of the results, i.e., to whom the results apply. We seek to provide the same for observational studies, with our algorithms producing an interpretable description of subjects with treatment group overlap.

Scenario 2: Evaluating guidelines. There are over 471 different guidelines for how to manage hypertension (Benavidez and Frakt, 2019). We could evaluate these—and new guidelines—using off-policy evaluation methods (Precup et al., 2000) on observational data derived from electronic medical records. Off-policy evaluation of a guideline is only possible on subsets of the population where there is some probability that the guideline was followed (which we will also call overlap). The estimated policy value should be accompanied by a description of the validity (overlap) region.

Beyond causal estimation, overlap is of interest in many other branches of machine learning: In domain adaptation, the overlap between source and target domains is the set of inputs for which we can expect a trained

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

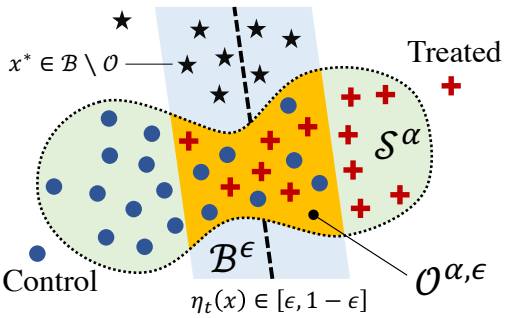

Figure 1: Overlap $\mathcal{O}^{\alpha,\epsilon}$ between treatment groups with joint support $\mathcal{S}^\alpha$. A point x^* has group propensity η_t bounded away from 0 and 1, but is outside of $\mathcal{O}^{\alpha,\epsilon}$.

model to transfer well (Ben-David et al., 2010; Johansson et al., 2019); In classification, overlap between inputs with different labels signifies regions that are hard to classify; In algorithmic fairness (Dwork et al., 2012), overlap between protected groups may shed light on disparate treatment of individuals from different groups who are otherwise comparable in task-relevant characteristics; In reinforcement learning, lack of overlap has been identified as a failure mode for deep Q-learning using experience replay (Fujimoto et al., 2019).

Our main contributions are as follows: (i) We propose desiderata in overlap estimation, and note how existing methods fail to satisfy them. (ii) We give a method for *interpretable characterization of distributional overlap*, which satisfies these desiderata, by reducing the problem to two binary classification problems, and using a linear programming relaxation of learning optimal Boolean rules. (iii) We give generalization bounds for rules minimizing empirical loss. (iv) We demonstrate that small rules often perform comparably to black-box estimators on a suite of real-world tasks. (v) We evaluate the interpretability of rules for describing treatment group overlap in post-surgical opioid prescription in a user study with medical professionals. (vi) We show how a generalized definition and method applies to policy evaluation and apply it to describing overlap in policies for antibiotic prescription.

2 RELATED WORK

Treatment group overlap is a central assumption in the estimation of causal effects from observational data. Comparing group-specific covariate bounds and lower-order moments is a common first step in assessing overlap (Rosenbaum, 2010; Zubizarreta, 2012; Fogarty et al., 2016) but fails to identify local regions of overlap when they exist (see the example of $\mathcal{O}^{\alpha,\epsilon}$ in Figure 1). An alternative is to estimate the *treatment propensity*—the probability that a subject was prescribed treatment. Treatment propensities bounded away from 0 and 1 at a point X indicates that treatment groups overlap at X (Rosenbaum and Rubin, 1983; Li et al., 2018).

In studies with partial overlap, it is common to restrict the study cohort by thresholding treatment propensity or discarding unmatched subjects after applying matching methods (Rosenbaum, 1989; Iacus et al., 2012; Kallus, 2016; Visconti and Zubizarreta, 2018). For example, Crump et al. (2009) proposed an optimal propensity threshold that minimizes the variance of the estimated average treatment effect on a sub-population. However, neither propensity thresholding nor matching are sufficient for guiding policy in new cases: they do not provide a self-contained, interpretable description of where treatment groups overlap *within* the study, nor do they provide insight into *external* validity by describing the limits of the study cohort.

Fogarty et al. (2016) address the first concern above by learning "interpretable study populations" through identifying the largest axis-aligned box that contains only subjects with bounded propensity. However, this approach is very limited in capacity and does not address external validity. For this reason, we strive to provide interpretable descriptions of overlap, both in terms of treatment propensity and the study support.

Rule-based models have been considered in classification tasks (Rivest, 1987; Angelino et al., 2017; Yang et al., 2017; Lakkaraju et al., 2016; Wang et al., 2017; Dash et al., 2018; Freitas, 2014; Wang and Rudin, 2015), subgroup discovery (Herrera et al., 2011) and density estimation (Ram and Gray, 2011; Goh and Rudin, 2015) but have to the best of our knowledge not been applied or tailored to support or overlap estimation.

3 DEFINING OVERLAP

We address *interpretable description of population overlap*. Our primary motivation is to aid *policy making* based on observational studies, the success of which relies on understanding and communicating the studies' validity region—the set of cases for which there is evidence that a particular policy decision is preferable. We identify the following desiderata for descriptions of overlap: (D.1) They cover regions where all populations (treatment groups) are well-represented; (D.2) They exclude all other regions, including those outside the support of the study (see Figure 1); (D.3) They can be expressed using a small set of simple rules. Next, we define overlap according to (D.1) and (D.2). We address (D.3) in Section 4.

Let subjects $i = 1, ..., m$ be observed through samples (x_i, t_i) of covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and a group indicator $T \in \mathcal{T}$. In our running example, X represents patient

attributes and T their treatment. We assume that subjects are independently and identically distributed according to a density $p(X,T)$, and that $\mathcal{X}$ is bounded. Let $p_t(X) := p(X \mid T = t)$ denote the covariate density of group $t \in \mathcal{T}$ and $\eta_t(x) := p(T = t \mid X = x)$ the propensity of membership in group $t \in \mathcal{T}$ for subjects with covariates $x \in \mathcal{X}$. We denote the probability mass of a set $S \subseteq \mathcal{X}$ under p by $P(S) := \int_{x \in S} dp$ and the support of p by $\text{supp}(p) := \{x \in \mathcal{X} : p(x) > 0\}$.

In the common case of two groups, $\mathcal{T} = \{0, 1\}$, overlap is typically defined as either a) the intersection of supports, $\text{supp}(p_0) \cap \text{supp}(p_1)$, or b) the set of covariate values for which all group propensities η_t are bounded away from zero (D'Amour et al., 2017; Li et al., 2018). We let $\mathcal{B}^\epsilon$ denote this latter set of values with *ϵ-bounded propensity* for a fixed parameter $\epsilon \in (0, 1)$ and an arbitrary set of groups $\mathcal{T}$,

$$\mathcal{B}^\epsilon := \{x \in \mathcal{X}; \forall t \in \mathcal{T} \colon \eta_t(x) > \epsilon\} \ . \tag{1}$$

Neither $\mathcal{B}^\epsilon$ nor the support intersection fully capture our desired notion of overlap: The former does not satisfy (D.2) since a point may have bounded propensity (true or estimated) but lie outside the population support $\text{supp}(p)$ (see Figure 1). Note that interpretable description alone does not address this. The latter is non-informative for variables with infinite support (e.g., a normal random variable), and even with finite support, we may wish to exclude distant outliers.

Our preferred definition of overlap combines the requirement of bounded propensity with a generalization of support called *α-minimum-volume sets* (Schölkopf et al., 2001). Let $\mathcal{C}$ be a set of measurable subsets of $\mathcal{X}$, let $V(C)$ denote the volume of a set $C \in \mathcal{C}$. An *α-minimum-volume* set $\mathcal{S}^\alpha$ of p is then

$$\mathcal{S}^\alpha := \arg\min_C \{V(C) \, ; P(C) \geq \alpha, C \in \mathcal{C}\} \ , \tag{2}$$

with $\mathcal{S}^1 = \text{supp}(p)$. For $\alpha < 1$, $\mathcal{S}^\alpha$ is not always unique, but the intersection S of two α-MV sets has mass $P(S) \geq 2\alpha - 1$. In this work, we let $\alpha < 1$ in order to handle distributions with infinite support and unwanted outliers, and *refer to $\mathcal{S}^\alpha$ as the support of p*. We define the *α, ϵ-overlap set*, for $\alpha, \epsilon \in (0, 1)$, to be

$$\mathcal{O}^{\alpha,\epsilon} := \mathcal{S}^\alpha \cap \mathcal{B}^\epsilon \ . \tag{3}$$

We define the problem of overlap estimation under definition (3) as *characterizing the set $\mathcal{O}^{\alpha,\epsilon}$ given thresholds α and ϵ*. In line with (D.3), these characterizations should be useful in policy making, and interpretable by domain experts, at small cost in accuracy. For notational convenience, we sometimes leave out superscripts from $\mathcal{S}^\alpha, \mathcal{B}^\epsilon$ and $\mathcal{O}^{\alpha,\epsilon}$, assuming that α, ϵ are fixed.

Remark. Defining overlap instead as the intersection of group-specific α-MV sets is feasible, but scales poorly with $|\mathcal{T}|$; it does not facilitate the generalization to policy evaluation described below; and the intersection of many descriptions may be hard to interpret.

3.1 Generalization to Policy Evaluation

The definition of $\mathcal{B}^\epsilon$ in (1) is motivated by causal effect estimation—comparison of outcomes under two or more alternative interventions. We may instead be interested in policy evaluation, which involves estimating the expected outcome under a conditional intervention π, which assigns a treatment t to each x following a conditional distribution $\pi(T|X)$ (Precup et al., 2000). To perform this evaluation, we only require that the propensity $p(T|X)$ of observed treatments be bounded away from zero for treatments which have non-zero probability under π. To describe the inputs for which this is satisfied, we generalize $\mathcal{B}^\epsilon$ to be a function of the target policy π,

$$\mathcal{B}^\epsilon(\pi) := \{x \in \mathcal{X}; \forall t : \pi(t \mid x) > 0 : \eta_t(x) > \epsilon\} \ . \tag{4}$$

More details are given in the supplement regarding the use of OverRule in this setting.

4 OVERRULE: BOOLEAN RULES FOR OVERLAP

We propose OverRule[1], an algorithm for identifying the overlap region $\mathcal{O}$ in (3) by first estimating the α-MV support set $\mathcal{S}$ (2) and then the bounded-propensity set $\mathcal{B}$ (1) restricted to $\mathcal{S}$, thereby satisfying desiderata (D.1)–(D.2). We aim to fulfill desideratum (D.3) by using Boolean rules—logical formulae in either disjunctive (DNF) or conjunctive (CNF) normal form—which have received renewed attention because of their interpretability (Dash et al., 2018; Su et al., 2016). See Figures 3–4 for examples of learned rules. OverRule proceeds in the following steps:

(i) Fit α-MV set $\hat{\mathcal{S}}^\alpha$ of $p(X)$ using Boolean rules

(ii) Fit model of group propensity $\hat{\eta}_{(\cdot)}$ over $\hat{\mathcal{S}}^\alpha$ and let $\tilde{b}(x) = \prod_{t \in \mathcal{T}} \mathbb{1}[\hat{\eta}_t(x) > \epsilon]$ define membership in $\tilde{\mathcal{B}}^\epsilon$

(iii) Approximate $\tilde{\mathcal{B}}^\epsilon$ using Boolean rules to yield $\hat{\mathcal{B}}^\epsilon$ and estimate overlap region by $\hat{\mathcal{O}}^{\alpha,\epsilon} = \hat{\mathcal{B}}^\epsilon \cap \hat{\mathcal{S}}^\alpha$.

In this section, we demonstrate how steps (i) & (iii) can be reduced to binary classification. This enables us to exploit the many existing methods for rule-based classification (Freitas, 2014) to improve the interpretability of $\hat{\mathcal{O}}$. Finally, we give results bounding the generalization error of estimates of both $\mathcal{S}$ and $\mathcal{S} \cap \mathcal{B}$.

[1]Code available at `https://github.com/clinicalml/overlap-code`

Remark. It was observed in evaluations with a medical practitioner that fitting rules for $\mathcal{S}$ and $\mathcal{B}$ separately improved interpretability as it makes clear which rules apply to which task and prevents the bulk of the rules from being consumed by one of the two tasks.

4.1 Estimation of S^α as Binary Classification

In the first step of OverRule, we learn a Boolean rule to approximate the α-MV set $\mathcal{S}^\alpha$ of the marginal distribution $p(X)$ by reducing the problem to binary classification between observed samples $\mathcal{D} \coloneqq \{x_i\}_{i=1}^m$ and uniform background samples. For clarity, we focus only on DNF rules—disjunctions of conjunctive clauses such as $(\text{Age} < 30 \wedge \text{Female}) \vee (\text{Married})$. As pointed out by Su et al. (2016), a CNF rule can be learned by swapping class labels and fitting a DNF rule.

We adapt previous notation and let $\mathcal{C}$ be a class of candidate α-MV sets $\mathcal{C}$ corresponding to Boolean rules, i.e., each $\mathcal{C}$ consists of the points in $\mathcal{X}$ that satisfy a rule. We will often not distinguish between a rule and its corresponding set $\mathcal{C}$ and thus will speak of the "volume" of a rule or clause. We aim to solve a normalized and regularized version of the α-MV problem in (2),

$$\underset{\mathcal{C}\in\mathcal{C}}{\arg\min}\ Q(\mathcal{C}) := \underbrace{\bar{V}(\mathcal{C})}_{\text{Volume}} + \underbrace{R(\mathcal{C})}_{\text{Regularization}} \quad \text{s.t.} \quad \underbrace{P(\mathcal{C}) \geq \alpha}_{\text{Coverage}} \tag{5}$$

where the volume $\bar{V}(\mathcal{C}) = V(\mathcal{C})/V(\mathcal{X}) \in [0,1]$ is normalized to that of $\mathcal{X}$. We assume that the regularization term $R(\mathcal{C})$ controls complexity by placing penalties λ_0 on each clause in the rule and λ_1 on each condition in a clause. Thus, for a Boolean rule with clauses $k = 1, \ldots, K$, each with p_k conditions, we have[2]

$$R(\mathcal{C}) = K\lambda_0 + \lambda_1 \sum_{k=1}^{K} p_k. \tag{6}$$

It is also assumed that the trivial "all-true" and "all-false" rules have complexity $R(\mathcal{C}) = 0$.

The volume $\bar{V}(\mathcal{C})$ may be difficult to compute repeatedly during optimization and $\mathcal{C}$ is often too large to allow pre-computation of $\bar{V}(\mathcal{C})$ for all $\mathcal{C}$. In particular, for DNF rules, each $\mathcal{C}$ is a union of potentially several overlapping clauses (see Figures 3–4 or the illustration in the supplement); even if the volume spanned by each clause is quick to compute on the fly, the overall volume may not be. As an alternative, the normalized volume $\bar{V}(\mathcal{C})$ can be estimated by means of uniform samples $\{x_{m+1}, \ldots, x_{m+n}\}$ over $\mathcal{X}$. Let $\mathcal{U}$ be the index set of these uniform samples. Then, $\frac{1}{n}\sum_{i\in\mathcal{U}} \mathbb{1}[x_i \in \mathcal{C}]$ is distributed as a scaled binomial random variable with mean $\bar{V}(\mathcal{C})$ and variance $\bar{V}(\mathcal{C})(1-\bar{V}(\mathcal{C}))/n$. Theorem 1 below provides guidance in selecting the number of uniform samples n to ensure a good estimate.

Given the above empirical estimator of volume, we reduce problem (5) to a classification problem between the marginal density $p(X)$ and a uniform distribution over $\mathcal{X}$. This reduction was also mentioned in the conclusion of Scott and Nowak (2006). We also replace the probability mass constraint with its empirical version over $\mathcal{D}$ with $\mathcal{I} = \{1, \ldots, m\}$. The result is a Neyman-Pearson-like classification problem with a false negative rate constraint of $1-\alpha$ (instead of the usual false positive constraint), as given below.

$$\begin{aligned} \hat{\mathcal{S}} := \underset{\mathcal{C}}{\arg\min} \quad & \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} \mathbb{1}[x_i \in \mathcal{C}] + R(\mathcal{C}) \\ \text{subject to} \quad & \sum_{i\in\mathcal{I}} \mathbb{1}[x_i \in \mathcal{C}] \geq \alpha m\ . \end{aligned} \tag{7}$$

The following theorem bounds the regret of the minimizer of (7) with respect to (5) and is proven in the supplement. The assumption of binary variables simplifies the analysis and is not a fundamental limitation.

Theorem 1. *Let $q^*(\alpha)$ denote the minimum regularized volume attained in* (5) *over the class of DNF rules with probability mass α. Assume that a) the regularization R follows* (6) *with fixed parameters λ_0, λ_1, b) all variables X_j are binary-valued, and c) the class $\mathcal{C}$ is restricted to rules satisfying necessary conditions of optimality for* (5) *(see Lemmas in the supplement). Then with probability greater than $1-2\delta$, the empirical estimate $\hat{\mathcal{S}}$ in* (7) *satisfies*

$$Q(\hat{\mathcal{S}}) \leq q^*(\alpha+\epsilon_m) + 2\epsilon_n \quad \textit{and} \quad P(\hat{\mathcal{S}}) \geq \alpha - \epsilon_m,$$

where $\epsilon_m = \sqrt{\frac{\lambda_1^{-1}\log(2d) + \left\lfloor 1+\log_2 \lambda_1^{-1}\right\rfloor \log \lambda_1^{-1} + \log(4/\delta)}{2m}}$ *and ϵ_n is defined analogously.*

Remark. The error term ϵ_m bounds the amount by which the probability constraint may be violated and contributes $q^*(\alpha+\epsilon_m) - q^*(\alpha)$ to the possible regret. Given the number of data samples m, penalty λ_1 (λ_0 does not appear in this simplified bound) could be chosen to keep ϵ_m small, although user preferences for rule complexity are likely to be more important in setting λ_0, λ_1. Given λ_1, the number of uniform samples n could in turn be chosen to reduce ϵ_n. Note that ϵ_m, ϵ_n are largely controlled by λ_1 and depend only logarithmically on the dimension d.

4.2 Estimation of $\mathcal{B}^\epsilon$ as Binary Classification

To estimate the set $\mathcal{B}^\epsilon$ of inputs with bounded group propensity $\eta_t(X) := p(T = t \mid X)$, we follow in the tradition of using black-box (potentially non-parametric)

[2] It is possible to generalize (6) to place different penalties on different conditions but we adopt (6) for simplicity.

estimators of propensity to identify overlapping or balanced cohorts in the study of causal effects (Crump et al., 2009; Fogarty et al., 2016). This is typically done by fitting a classifier (e.g., logistic regression) for predicting T given X, and letting $\hat{\eta}_t(x)$ be the estimated probability of class t for input x. Given such an estimate, we assign a label $\tilde{b}_i$ to each data point $x_i \in \mathcal{D}$ indicating significant propensity for every group,

$$\forall i \in [m] : \tilde{b}_i = \prod_{t \in \mathcal{T}} \mathbb{1}[\hat{\eta}_t(x_i) \geq \epsilon] \ . \tag{8}$$

Let $\tilde{\mathcal{B}} = \{x_i : \tilde{b}_i = 1\}$. Similar to the case of $\mathcal{S}^\alpha$, we may now reduce estimation of $\mathcal{B}^\epsilon$ to binary classification. Given $\hat{\mathcal{S}}$, the minimizer of (7), we again set up a Neyman-Pearson-like classification problem, now regarding the intersection $\hat{\mathcal{S}} \cap \tilde{\mathcal{B}}$ as the positive class:

$$\begin{aligned} \hat{\mathcal{B}} := \underset{C}{\arg\min} \ & \frac{1}{|\hat{\mathcal{S}} \setminus \tilde{\mathcal{B}}|} \sum_{i : x_i \in \hat{\mathcal{S}} \setminus \tilde{\mathcal{B}}} \mathbb{1}[x_i \in \mathcal{C}] + R(\mathcal{C}) \\ \text{subject to} \ & \sum_{i : x_i \in \hat{\mathcal{S}} \cap \tilde{\mathcal{B}}} \mathbb{1}[x_i \in \mathcal{C}] \geq \beta |\hat{\mathcal{S}} \cap \tilde{\mathcal{B}}| \ . \end{aligned} \tag{9}$$

The sets $\hat{\mathcal{S}} \setminus \tilde{\mathcal{B}}$ and $\hat{\mathcal{S}} \cap \tilde{\mathcal{B}}$ are defined by the solution to (7) and the base estimator (8). To accommodate the policy evaluation setting described in Section 3, we can modify the pseudo-labels defined in (8) to be $\tilde{b}_i(\pi) = \prod_{t \in \pi(x_i)} \mathbb{1}[\hat{p}(T = t \mid X = x_i) \geq \epsilon]$, where $\pi(x_i) \coloneqq \{t : \pi(t|x_i) > 0\}$, and solve (9) using $\tilde{\mathcal{B}}(\pi) = \{x_i : \tilde{b}_i(\pi) = 1\}$ in place of $\tilde{\mathcal{B}}$. The resulting full procedure is given in the supplement.

Generalization of the final estimator. In the supplement, we state and prove a theorem bounding the generalization error of our final estimator, $\hat{\mathcal{O}} = \hat{\mathcal{S}} \cap \hat{\mathcal{B}}$. It shows that for good base estimators $\hat{\mathcal{S}}, \tilde{\mathcal{B}}$, the error of $\hat{\mathcal{O}}$ with respect to the true overlap $\mathcal{O}$ is dominated by its error with respect to the base estimators. Hence, practitioners may make an informed tradeoff between accuracy and interpretability based on this metric.

4.3 Optimizing Boolean Rules

Next, we describe a procedure for optimizing (7) over a class $\mathcal{C}$ of Boolean DNF rules. The same procedure also solves (9).

We assume that base features X have been binarized to form literals such as (Age > 30) or (Sex $=$ Female), as is standard in e.g. decision tree learning. A conjunction may thus be represented as the product of binary indicators of these literals. We let $\mathcal{K}$ index the set of all possible (exponentially many) conjunctions of literals, e.g. (Age > 30) $\wedge$ Female. Then, for $k \in \mathcal{K}$, let $a_{ik} \in \{0, 1\}$ denote the value taken by the k-th conjunction at sample x_i. Let the DNF rule be parametrized by $r \in \{0, 1\}^{|\mathcal{K}|}$ such that $r_k = 1$ indicates that the k-th conjunction is used in the rule.

Define an error variable ξ_i for i in $\mathcal{U} \cup \mathcal{I}$ representing the penalty for covering or failing to cover point i, depending on its set membership. Then, problem (7) may be reformulated as follows,

$$\begin{aligned} \underset{r}{\text{minimize}} \quad & \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \xi_i + R(r) \\ \text{subject to} \quad & \begin{cases} r_k \in \{0, 1\}, \ k \in \mathcal{K}, \\ \xi_i \geq 1 - \sum_{k \in \mathcal{K}} a_{ik} r_k, \ \ \xi_i \geq 0, \ i \in \mathcal{I}, \\ \sum_{i \in \mathcal{I}} \xi_i \leq (1 - \alpha) m \\ \xi_i = \max_{k \in \mathcal{K}} (a_{ik} r_k), \ i \in \mathcal{U}. \end{cases} \end{aligned} \tag{10}$$

Problem (10) is an IP with an exponential number of variables and is intractable as written. We follow the column generation approach of Dash et al. (2018) to effectively manage the large number of variables and solve (10) approximately. As in that previous work, we bound from above the max in the last constraint of (10) with the sum (Hamming loss instead of zero-one loss) as it gives better numerical results. The choice of regularization in (6) implies $R(r) = \sum_{k \in \mathcal{K}} \lambda_k r_k$ with $\lambda_k = \lambda_0 + \lambda_1 p_k$. Thus the objective becomes linear in r, $\sum_{k \in \mathcal{K}} \left(1/|\mathcal{U}| \sum_{i \in \mathcal{U}} a_{ik} + \lambda_k\right) r_k$, and the ξ_i, $i \in \mathcal{U}$ constraints are absorbed into the objective. We then follow the overall procedure in (Dash et al., 2018) of solving the linear programming (LP) relaxation, using column generation to add variables only as needed.

We make the following departures from Dash et al. (2018). As noted, (10) has a constraint on false negative rate instead of a corresponding objective term and a complexity penalty $R(r)$ while Dash et al. (2018) use a constraint. As a result, the LP reduced costs, needed for column generation, are different. With dual variables $\mu_i \geq 0$, $i \in \mathcal{I}$ corresponding to the ξ_i, $i \in \mathcal{I}$ constraints in (10), the reduced cost of conjunction k is now $1/|\mathcal{U}| \sum_{i \in \mathcal{U}} a_{ik} + \lambda_k - \sum_{i \in \mathcal{I}} \mu_i a_{ik}$, which remains a linear function of a_{ik}, allowing the same column generation method to be used. We also avoid the need for an IP solver as used in Dash et al. (2018) by a) solving the column generation problem using a beam search algorithm from (Wei et al., 2019), and b) restricting (10) to the final columns once column generation terminates, converting to a weighted set cover problem, and applying a greedy algorithm to obtain an integer solution.

5 EXPERIMENTS

In our experiments, we seek to address the following questions, while relating the performance of OverRule

to that of MaxBox (MB) (Fogarty et al., 2016), which is also designed to produce interpretable study populations. (i) **Why is support estimation important?** In Section 5.1 we give a conceptual illustration using the Iris dataset, where MaxBox returns a description that empirically includes a large space outside of the true overlap region. (ii) **How well does OverRule approximate the base estimators / true overlap region?** In Section 5.2 we use the Jobs (LaLonde, 1986) dataset to show that performance of OverRule is comparable to that of the base estimators, and generally surpasses the performance of MaxBox. (iii) **Do the resulting rules yield any insights?** We apply OverRule to overlap estimation in two real-world clinical datasets on (1) post-surgical opioid prescriptions, and (2) policy evaluation in antibiotic prescriptions. For the former, we conducted a user study with three clinicians to interpret and critique the output, with additional comparison to the output of MaxBox.

OverRule and MaxBox algorithms are both *meta-algorithms* in the sense that they take (as input) labels indicating whether each data point is in the overlap set. To generate these labels, we use a variety of base overlap estimators: (i) *Covariate Bounding Boxes*: The intersection of covariate (marginal) bounding boxes (CBB), analogous to classical balance checks in causal inference. The bounding boxes are selected to cover the $[(1-\alpha)/2, (1+\alpha)/2]$ quantiles of the data. (ii) *Propensity Score Estimators*: Standard propensity score estimators as described in (8) and Crump et al. (2009) with logistic regression (PS-LR) or k-nearest neighbors (PS-kNN) estimates of the propensity. These can be viewed as a binary version of overlap weights (Li et al., 2018). (iii) *One-Class SVMs*: One-Class Support Vector Machines (OSVM) to first estimate conditional supports and then use their intersection as overlap labels. Details on hyperparameter selection and feature binarization are given in the supplement, along with general guidance on hyperparameter selection depending on user goals, from optimizing an observable metric (e.g., accuracy w.r.t the base estimator), to generating shorter rule sets, to exploring structure in the data.

5.1 Illustrative Example: Iris

We use the Iris dataset to illustrate the importance of combining explicit support estimation (lacking in MaxBox) with an interpretable characterization of the overlap region (lacking in propensity score models). We use OverRule to identify the overlap between members of two species of Iris, as represented by their sepal and petal dimensions. In Figure 2, we visualize the estimates $\hat{\mathcal{O}}$ learned using OverRule and MaxBox in the space of sepal length and width. In contrast, the coefficients of a logistic regression propensity score model, $[-1.7, -1.5, 2.5, 2.6]^\top$ reveal very little about which points lie in the overlap set.

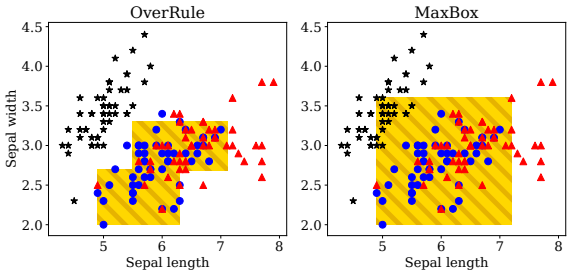

Figure 2: Overlap (orange stripes) between Versicolor (blue circles) and Virginica (red triangles) species in the Iris dataset as identified by OverRule (left) and MaxBox (right) using the same base estimator of propensity. Black stars indicate samples of the (unobserved) Setosa species. We see that MaxBox identifies several of the Setosa samples as being in the overlap set, despite it being outside of the support of the observed data.

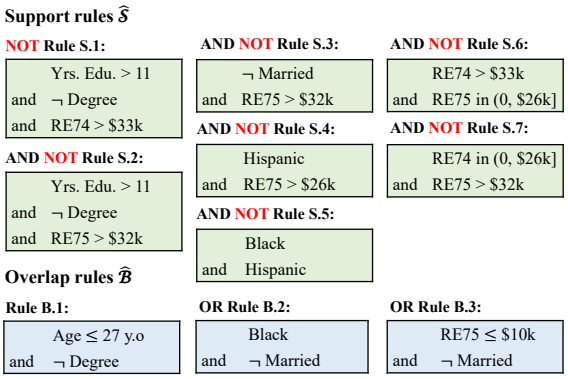

Figure 3: OverRule description of the overlap region $\mathcal{O}$ in the Jobs dataset learned using the LR propensity base estimator, achieving held-out balanced accuracy of 0.88. ¬ indicates a negation, and CNF support rules are given with rule-level negations applied for readability. If *none* of the support rules (top) and *any* of the overlap rules (bottom) apply, a subject is in $\mathcal{O}$.

5.2 Job Training Programs

In this section, we demonstrate that OverRule compares favorably to MaxBox in terms of approximating both the derived overlap labels (using a base estimator), as well as the "ground truth" overlap labels in a real dataset. To do so, we use data from a famous trial performed to study the effects of job training (LaLonde, 1986; Smith and Todd, 2005), in which eligible US citizens were randomly selected into ($T = 1$), or left out of ($T = 0$) job training programs. The RCT ($E = 1$), which satisfies overlap by definition, has since been combined with non-experimental control samples

Table 1: Overlap estimation in Jobs. Balanced accuracy (Acc), false positive rate (FPR), false negative rate (FNR), and number of literals (L) with standard deviations over 5-fold CV. MB and OR indicate MaxBox and OverRule. MB did not run with CBB.

	Acc	FPR	FNR	L
Baselines (base estimators):				
CBB	0.75 ± 0.02	0.12 ± 0.01	0.38 ± 0.03	—
OSVM	0.82 ± 0.01	0.22 ± 0.03	0.14 ± 0.02	—
PS-k-NN	0.90 ± 0.02	0.14 ± 0.02	0.05 ± 0.02	—
PS-LR	0.96 ± 0.01	0.10 ± 0.01	0.09 ± 0.03	—
MaxBox with base estimator:				
OSVM	0.68 ± 0.01	0.09 ± 0.02	0.54 ± 0.01	16
PS-kNN	0.84 ± 0.01	0.03 ± 0.01	0.29 ± 0.02	16
PS-LR	0.80 ± 0.02	0.04 ± 0.01	0.35 ± 0.04	16
OverRule with base estimator:				
CBB	0.83 ± 0.01	0.16 ± 0.01	0.19 ± 0.02	20
OSVM	0.84 ± 0.02	0.25 ± 0.03	0.07 ± 0.02	23
PS-kNN	0.89 ± 0.02	0.16 ± 0.02	0.06 ± 0.02	40
PS-LR	0.88 ± 0.02	0.15 ± 0.04	0.09 ± 0.01	21

$(E = 0, T = 0)$, forming a larger observational set (Jobs), to serve as a benchmark for causal effect estimation (LaLonde, 1986). Here, we aim to characterize the overlap between treated and control subjects.

Due to the trial's eligibility criteria, the experimental and non-experimental cohorts barely overlap; standard logistic regression separates the experimental and non-experimental groups with held-out balanced accuracy of 0.96. Since all treated subjects were part of the experiment, the experimental cohort perfectly represents the overlap region. For this reason, we use the experiment indicator E as ground truth for $\mathcal{O}$, at the risk of introducing a small number of false negatives. In studies of causal effects in this data, the following features were included to adjust for confounding: Age, #Years of education (Educ), Race (black/hispanic/other), Married, No degree (NoDegr), Real earnings in 1974 (RE74) and 1975 (RE75). These are the features X for which we estimate overlap.

We present results in Table 1 and Figure 3, where all balanced accuracies are w.r.t. the ground truth indicator E. For the propensity base estimators, the OverRule approximations achieve slightly lower balanced accuracies than the base estimator, but with a simpler description, while for the other base estimators the accuracy is actually better. OverRule compares favorably to MaxBox on balanced accuracy, although MaxBox generally achieves a lower FPR, likely because it does not try to retain a fixed fraction β of the overlap set. In the supplement, we show that the held-out balanced accuracy quickly converges as the number of literals in the rules increases and correlates strongly with the quality by which the rule set approximates the base estimator.

The learned support rules in Figure 3 demonstrate that support estimation can find gaps in the dataset that are intuitive, such as a lack of individuals with high income but no degree (Rules S.1-2) or whose income changes dramatically from 1974 to 1975 (Rules S.6-7). The learned overlap rules conform to expectations, as the eligibility criteria for the RCT allow only subjects who were currently unemployed and had been so for most of the time leading up to the trial—factors that correlate with age and education (Rule B.1), previous income (Rule B.3), and marital status (Rules B.2-3) (Smith and Todd, 2005).

5.3 Post-surgical Opioid Prescriptions

Opioid addiction affects millions of Americans. Understanding the factors that influence the risk of addiction is thus of great importance. To this end, Brat et al. (2018) and Zhang et al. (2017) study the effect of choices in opioid prescriptions on the risk of future misuse. Here, we study a group of *post-surgical* patients who were given opioid prescriptions within 7 days of surgery, replicating the cohort eligibility criteria of Brat et al. (2018) using a subset of the MarketScan insurance claims database. We compare groups of patients with morphine milligram equivalent (MME) doses above and below the 85th percentile in the cohort, MME=450. Subjects were represented by basic demographics (age, sex), diagnosis history, and procedures billed as surgical on the index date (not mutually exclusive). Cohort statistics are given in the supplement. We fit three models: An OverRule model (OR) using DNF support rules and a random forest base estimator, a MaxBox model (MB) (Fogarty et al., 2016) with the same base estimator, and another OverRule model describing the complement of $\mathcal{O}$ (OR-C). The balanced accuracies of these models w.r.t. the base were 0.90 (OR), 0.77 (MB) and 0.92 (OR-C). Learning took 10 minutes for OverRule (Python) and 7 minutes for MaxBox (R). Other hyperparameter details are in the supplement.

In Figure 4, we summarize the rules learned by OR which cover 27% of the overall population. MB learned: (Musculoskeletal surg. $\wedge$ ¬Mediastinum surg. $\wedge$ ¬Male genital surg. $\wedge$ ¬Maternity surg. $\wedge$ ¬Lumbosacral spondylosis without myelopathy) which covers 17% of patients. The rules learned by OR-C are presented in the supplement.

To evaluate the interpretability of the output, we conducted a qualitative user study through a moderated discussion with three participants: two attending surgeons (P1 & P2) and a 4th year medical student (P3) at a large US teaching hospital. Before seeing the outputs of any method, the participants were asked to give their expectations for what to find in the overlap set.

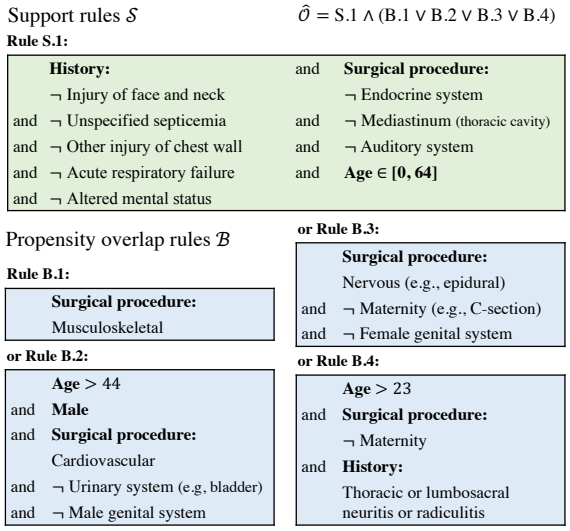

Figure 4: OverRule description of post-surgical patients likely to receive both high and low opioid doses. A patient is in the overlap set if the support rule (top) applies and *any* propensity overlap rule (bottom) applies. ¬ indicates negation. The rules cover 27% of patients with balanced accuracy of 0.90 w.r.t. the base estimator. Surgical procedures are not mutually exclusive.

The participants expected that the overlap set would mostly correspond to patients in the higher dose range, as these patients are often considered also for smaller doses, and that overlap would be driven largely by surgery type. All participants expected Musculoskeletal and Cardiovascular surgery patients to be predominantly in the higher dose group, and sometimes in the lower, and one suggested that Maternity surgeries (e.g., C-sections) would be only in the lower range. These comments are all consistent with the findings of OverRule, which identified all of these surgery types as important. MaxBox identified only Musculoskeletal surgery patients as overlapping. One participant expected history of psychiatric disease and Tobacco use disorder to be predictive of higher prescription doses for some patients, and thus overlap. Neither method identified psychiatric disease, but Tobacco use disorder was identified by OR-C as predictive (see supplement).

The participants found the support rules ($\hat{\mathcal{S}}$) output by OR (Figure 4 top) intuitive. P1 stated that Endocrine surgeries are not typically followed by opioid prescriptions. They found the MaxBox and OR rule descriptions easy to interpret, and discussion focused on their clinical meaning. The first three propensity overlap rules B.1-B.3 were all consistent with expectation as described above, with the caveat that Cardiovascular patients are not typically stratified by Urinary and Genital surgeries. This was later partially explained by catheters being billed as Urinary and P3 interpreted this as a proxy for more severe Cardiovascular surgeries. P1 pointed out the value in discovering such surprising patterns that may be hidden in black-box analyses. The OR-C rules were found hard to interpret due to many double negatives ("excluded from exclusion"), but were ultimately deemed clinically sound.

Remark: We noted that these support rules primarily exclude individually rare features, in lieu of e.g., finding that certain non-rare surgery types do not co-occur. This motivated both (1) an empirical study (w/semi-synthetic data) of how support rule hyperparameters influence the recovery of these interactions, and (2) the generation of new rules. Both are in the supplement.

5.4 Policy Evaluation of Antibiotic Prescription Guidelines

Using the policy evaluation formulation of $\mathcal{B}^{\epsilon}(\pi)$ (Section 3.1), we apply OverRule to assess overlap for a policy that follows clinical guidelines published by the Infectious Disease Society of America (IDSA) for treatment of uncomplicated urinary tract infections (UTIs) in female patients (Gupta et al., 2011). Using medical records from two academic medical centers, we apply OverRule to a cohort of 65,000 UTI patients to test whether it can recover a clinically meaningful overlap set. From a qualitative perspective, we discussed the resulting rules with an infectious disease specialist, who verified that they have a clear clinical interpretation as identifying primarily outpatient cases and uncomplicated inpatient cases, which are where the guidelines are applied in practice. Detailed results (including quantitative results) are given in the supplement.

6 CONCLUSION

We have presented OverRule—an algorithm for learning rule-based characterizations of overlap between populations, or the inputs for which policy evaluation from observational data is feasible. The algorithm learns to exclude points that are marginally out-of-distribution, as well as points where some population/policy has low density. We gave theoretical guarantees for the generalization of our procedure and evaluated the algorithm on the task of characterizing overlap in observational studies. These results demonstrated that our rule descriptions often have similar accuracy to black-box estimators and outperform a competitive baseline. In an application to study treatment-group overlap in post-surgical opioid prescription, a qualitative user study found the results interpretable and clinically meaningful. Similar observations were made in an application to evaluation of antibiotic prescription policies. Future research challenges include investigating the scalability of the method with the dimensionality of the input.

Acknowledgments

We thank Chloe O'Connell and Charles S. Parsons for providing clinical feedback on the opioid misuse experiment, and Sanjat Kanjilal for providing clinical feedback on the antibiotic prescription experiment. We also thank Bhanukiran Vinzamuri for assistance with the opioids data, David Amirault for insightful suggestions and feedback, and members of the Clinical Machine Learning group for feedback on earlier drafts. This work was partially supported by the MIT-IBM Watson AI Lab and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 35–44.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Benavidez, G. and Frakt, A. B. (2019). Fixing clinical practice guidelines. Health Affairs Blog, August 5, Retrieved from: `https://www.healthaffairs.org/do/10.1377/hblog20190730.874541/full/`.

Brat, G. A., Agniel, D., Beam, A., Yorkgitis, B., Bicket, M., Homer, M., Fox, K. P., Knecht, D. B., McMahill-Walraven, C. N., Palmer, N., et al. (2018). Post-surgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study. *Bmj*, 360:j5790.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2017). Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

Dash, S., Gunluk, O., and Wei, D. (2018). Boolean decision rules via column generation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 4660–4670. Curran Associates, Inc.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F., and Small, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514):447–458.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10.

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2052–2062. PMLR.

Goh, S. T. and Rudin, C. (2015). Cascaded high dimensional histograms: A generative approach to density estimation. *arXiv preprint arXiv:1510.06779*.

Gupta, K., Hooton, T. M., Naber, K. G., Wullt, B., Colgan, R., Miller, L. G., Moran, G. J., Nicolle, L. E., Raz, R., Schaeffer, A. J., and Soper, D. E. (2011). International clinical practice guidelines for the treatment of acute uncomplicated cystitis and pyelonephritis in women: A 2010 update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. *Clinical Infectious Diseases*, 52(5):e103–20.

Herrera, F., Carmona, C. J., González, P., and Del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Johansson, F., Sontag, D., and Ranganath, R. (2019). Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684. ACM.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

National Cancer Institute (2012). Bortezomib in treating patients with newly diagnosed multiple myeloma. ClinicalTrials.gov Identifier NCT00075881. Retrieved from: `https://clinicaltrials.gov/ct2/show/NCT00075881`.

Precup, D., Sutton, R. S., and Singh, S. P. (2000). Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 759–766.

Ram, P. and Gray, A. G. (2011). Density estimation trees. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–635. ACM.

Rivest, R. L. (1987). Learning decision lists. *Machine learning*, 2(3):229–246.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenbaum, P. R. (2010). *Design of observational studies*, volume 10. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Scott, C. D. and Nowak, R. D. (2006). Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353.

Su, G., Wei, D., Varshney, K. R., and Malioutov, D. M. (2016). Learning sparse two-level Boolean rules. In *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pages 1–6.

Visconti, G. and Zubizarreta, J. R. (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4:217–249.

Wang, F. and Rudin, C. (2015). Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and MacNeille, P. (2017). A Bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18(70):1–37.

Wei, D., Dash, S., Gao, T., and Gunluk, O. (2019). Generalized linear rule models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Yang, H., Rudin, C., and Seltzer, M. (2017). Scalable Bayesian rule lists. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 1013–1022.

Zhang, J., Iyengar, V., Wei, D., Vinzamuri, B., Bastani, H. S., Macalalad, A. R., Fischer, A. E., Yuen-Reed, G., Mojsilovic, A., and Varshney, K. R. (2017). Exploring the causal relationships between initial opioid prescriptions and outcomes. In *AMIA Workshop on Data Mining for Medical Informatics*, Washington, DC.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.