Supplementary Material for 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

The supplement is structured as follows:

- **Guidance on hyperparameter selection**: We take a deeper dive into the impact of hyperparameter selection on support and overlap estimation, including an in-depth empirical evaluation with concrete recommendations on how to set hyperparameters for support estimation given an a-priori belief that higher-order intersections of variables may be excluded from the cohort.
- **Application to Policy Evaluation**: We discuss in more depth how the OverRule algorithm can be applied to finding areas of sufficient coverage for policy evaluation tasks.
- **Additional experimental results**: In addition to providing additional detail on the experiments presented in the main paper, we also present several results that were only alluded to in the main paper. This includes the detailed results for the policy evaluation task (antibiotic prescription), as well as additional rules learned for the opioids prescription task.
- **Theoretical results**: We include proofs for our theoretical results, as well as an additional Theorem bounding the generalization error of our two-stage estimator in terms of the error of the base estimators.

In addition, to build further intuition for Boolean rules, we illustrate a Boolean rule in the DNF form in a 2D example in Figure S1.

Code for this paper can be found at `https://github.com/clinicalml/overlap-code`

A Choosing Hyperparameters

A.1 Overview

Considering OverRule along with the base estimator, there are a few distinct sets of hyperparameters to choose

- **Support Rules**: The support rule estimation task requires a specification of DNF versus CNF form, a specification of $\alpha, \lambda_0, \lambda_1$ used in the objective, and the number of samples to draw from the reference measure.
- **Base Estimator and Overlap Labels**: In addition to the hyperparameters of the base estimator itself, a threshold ϵ must be chosen to generate overlap labels
- **Overlap Rules**: These rules similarly require a specification of DNF or CNF form, and specification of $\beta, \lambda_0, \lambda_1$.

For the base estimator itself, the hyperparameters can be tuned in the usual way using cross-validation using a metric of interest (e.g., AUC). The choice of ϵ is studied in the existing literature (Crump et al., 2009) and ultimately depends on the downstream causal inference task, though $\epsilon = 0.1$ is sometimes considered as a rule of thumb. For the support rules, we typically set the number of reference measure samples to be as large as computationally feasible.

For the overlap and support rules, the remaining hyperparameters can be chosen (1) by using cross-validation to optimize for balanced accuracy (or some other metric, like false positive rates) with respect to the overlap labels or uniform background samples, (2) with some other objective in mind, e.g., setting the λ parameters

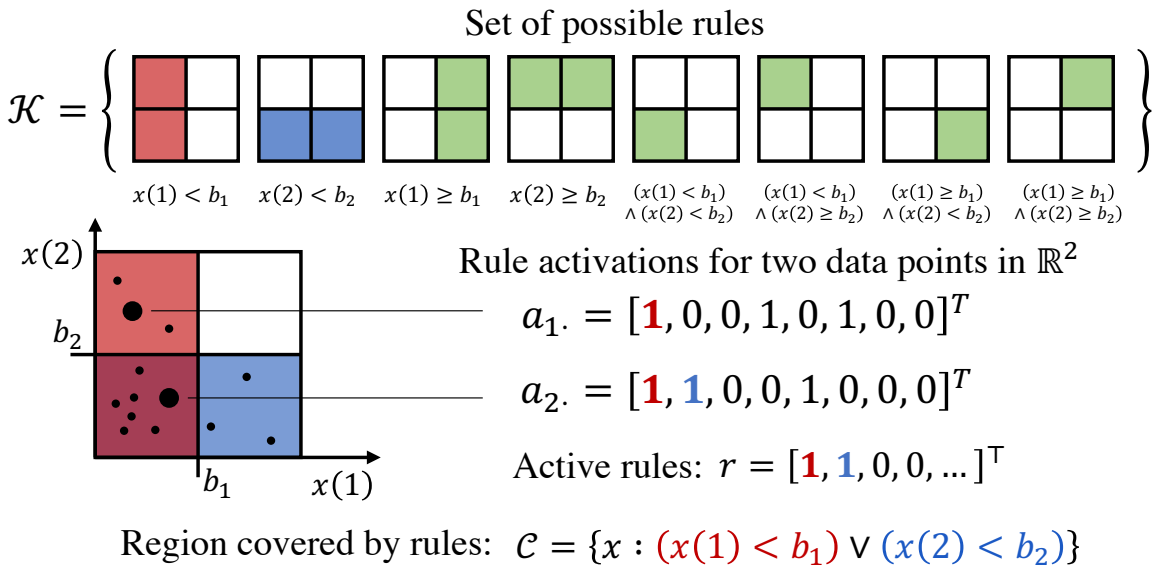

Figure S1: Boolean rules on disjunctive normal form (DNF). We highlight data points represented by their activations, $a_{1\cdot}, a_{2\cdot}$ of rules from the set $\mathcal{K}$ of all possible rules. $\mathcal{C}$ is the region described by the rule set and r indicators for the rules.

to be large to discourage many rules, even if more rules would increase accuracy, or (3) with the goal in mind of choosing values (or exploring a range of values) most likely to discover "interesting patterns" in the cohort.

We expand upon a concrete instance of this latter goal in the remainder of this section, particularly as regards hyperparameter selection for support estimation, where extremely high accuracy is particularly easy to achieve and is thus less informative for the purposes of hyperparameter selection.

A.2 Choosing Support Hyperparameters to highlight exclusions

Motivation: In the context of our motivating applications, the primary purpose of support estimation is to identify regions where we do not have any (or have very few) observations. For instance, if there are no men in our dataset who also have cardiac arrhythmia[1], then this would be a clinically relevant fact that should be highlighted. Thus, we would like to select hyperparameters which minimize our risk of overlooking these types of exclusions.

In this section we give some guidance on how to select hyper-parameters for support estimation with this particular goal in mind, based on synthetic and real-data experiments. To recap, these hyper-parameters include (i) α, the support level, and (ii) λ_0, λ_1, regularization parameters for learning support rules. There are also relevant hyperparameters in the underlying algorithm of Wei et al. (2019), primarily the width of the beam search used during column generation.

Summary: For this purpose, we recommend setting $\alpha \approx 1$, and in particular we consistently observed best results for $\alpha \geq 0.98$. We observe that for α sufficiently close to 1, the results are less sensitive to different values of λ_0, λ_1. In addition, we recommend setting the width of the beam search in the algorithm of Wei et al. (2019) to be on the same order of magnitude as the number of binary features.

These recommendations have the effect of encouraging the algorithm to consider higher-order interactions between variables that describe regions with little or no support in the data (e.g., "there are no men with cardiac arrhythmia"), and we verify this through experiments where we selectively remove regions of the data, and verify whether or not the algorithm can recover these regions.

Concretely, we use both a synthetic and semi-synthetic case where we manually exclude all points which satisfy a simple boolean rule, and look to identify that exclusion automatically. That is to say, in both cases we take a dataset and **remove** data points $\mathbf{x} \in \{0, 1\}^d$ which satisfy a rule of the form $x_i = 1 \wedge x_j = 1$ for two features x_i, x_j, and then check if our algorithm incorporates this into the learned rule set.

- **Synthetic Case:** In this setting, we generate data comprised of 22 independent binary features, such that 10 features are rare (binomial with $p = 0.01$),

[1]This would be surprising, as men with arrhythmia are fairly common in the general population

12 features are common ($p = 0.5$), and we remove all data points which satisfy a conjunction of the last two common features.

- **Semi-Synthetic Case (Antibiotic Prescription):** In this setting, we used the medical records dataset described in Section 5.4, and removed all men with cardiac arrhythmia, which compromised 5% of the total population.

This particular type of exclusion benefits from a CNF formulation (AND of ORs) of the support task. This is because the exclusion can be described in a parsimonious way (independently of other aspects of support) as a single additional rule. As discussed in Section 4.1, it is straightforward to convert the CNF formulation to a DNF formulation and vice versa. However, we note that the CNF formulation (for a fixed number of reference samples) can be more computationally intensive than the DNF formulation.

A.2.1 Synthetic Experiments

For the synthetic case, our goal is to build intuition that we can validate in the semi-synthetic setting. We will first describe our data-generating process in more detail, and then describe the results and conclusions from an exhaustive hyperparameter search.

Synthetic Data Generation: We generate data as follows. Note that we are only concerned (for the moment) with estimating support, so we do not include any notion of treatment groups.

- We sample 10,000 data points $x \in \{0,1\}^d$ where $d = 22$, by sampling (for each data point):
  - 10 "rare" binary features $r_1, \ldots, r_{10}$, generated independently with $p = 0.01$
  - 12 "common" binary features $c_1, \ldots, c_{12}$, generated independently with $p = 0.01$
  - Thus, each data point is given by $\mathbf{x} = [r_1, \ldots, r_{10}, c_1, \ldots, c_{12}]$
- We remove all data points which satisfy $c_{11} = 1 \wedge c_{12} = 1$, which is approximately 25% of all data points. Our goal is to recover the corresponding **inclusion rule** as part of the final rule set of $c_{11} = 0 \vee c_{12} = 0$.

Hyperparameter Search & Outcomes: With this setup, we estimate support using the algorithm given in the main paper, using every combination of the following hyperparameters

- $\alpha \in \{0.95, 0.96, 0.97, 0.98, 0.99\}$, the constraint on covering our data.
- $\lambda_0 \in \{0, 10^{-6}, 10^{-4}, 10^{-2}\}$, and $\lambda_1 \in \{10^{-6}, 10^{-4}, 10^{-2}\}$, the regularization terms.
- $B \in \{10, 15, 20, 25, 30\}$, the width of the beam search used in Wei et al. (2019)

For each combination of hyperparameters, we run the experiment three times, generating a new set of fake data with each run. The same three random seeds are used across all hyperparameter combinations. We recorded a number of relevant outcomes, including

- Does the final rule set include the inclusion rule $c_{11} = 0 \vee c_{12} = 0$?
- How many rules are considered in the final rule set, and how long (on average) are these rules?
- How many "perfect" rules are found, which exclude none of the generated data points?

Observations: The full results of the hyperparameter search are given in Table S5, but we summarize our observations and recommendations here.

- *Recovery by LP → recovery by rounded rules:* Across all hyperparameter settings, if the desired inclusion rule was found during column generation (and thus considered by the LP), it was uniformly included in the final rounded rule.[2] Thus, our goal is to ensure that the desired inclusion is picked up by the LP during column generation.
- *Beam Search Width should be higher than # features:* Recall that the LP relaxation with column generation starts by considering only rules with a single literal, and beam search is used to select additional rules for consideration, with a maximum width of B. If B is lower than the number of rare features, then the first B rules considered will tend to be rules on single rare features. This prevents the beam search from exploring interactions between more prevalent features. Setting the beam-search width to a sufficiently high number ($\approx$ total features) forces the column generation to explore all rules with two literals, helpful for recovery of our desired inclusion rule. This is demonstrated in Table S1.
- *Higher values of α produce more stable results across λ.* Higher values of α tends to render the results less sensitive to choice of regularization λ, and tends to produce more reliable results in terms of recovery of our desired rule. As demonstrated in

[2] This is not a general rule; While it holds in the synthetic case, it will not hold exactly in the semi-synthetic case with real data, as demonstrated in the next section.

Tables (S2a-S2c), lower values of α are more sensitive to λ_1 in terms of both recovering the desired exclusion, as well as the number of rules found. At higher values of α, there is more consistent recovery of "perfect" rules, which exclude none of the sample points (and hence do not contribute to the constraint).

Table S1: Beam Search Width and proportion of runs (across all other hyperparameter settings of $\alpha, \lambda_0, \lambda_1$) in which the synthetic region was correctly identified by the final rule set ("Rounded"). Once the beam search width is sufficiently high (larger than the number of rare features), further increasing it does not appear to help.

Beam Width	10	15	20	25	30
Recovered	0.07	0.87	0.87	0.87	0.87

Discussion / Intuition: Due to the greedy nature of the column generation procedure, a common failure mode is to only consider rules that include rare features, because those singleton rules exclude a significant amount of reference measure, and excluding rare features does not violate the α-constraint. For instance, a support rule of the form "not one of these K rare features" will (roughly speaking) exclude K percent of the samples (if each rare feature has 1% prevalence), while producing a volume of 2^{-K}. Thus, an overly greedy approach can obtain an objective value that is exponentially small in the number of rare features excluded, as long as it does not hit the α constraint. This has the effect of "crowding out" more complex rules.

Take a concrete example in Table S2b to build intuition for how the greedy set covering algorithm can fail in this case: Suppose $\lambda_0 = 0$, $\lambda_1 = 0.01$, and $\alpha = 0.95$, and suppose that our current solution excludes 5 rare features before hitting the α constraint, then the reference volume is given by $2^{-5} \approx 0.03$. In this case, adding the desired inclusion rule will reduce the volume by 1/4 (a reduction in absolute terms which is < 0.01) while increasing the regularization penalty by 0.02. Thus, it will not be included.

To avoid this failure mode, we can increase α, which has the effect of reducing the number of singleton rules K that can be added before violating the constraint.

A.2.2 Semi-Synthetic Experiments

In the semi-synthetic experiment, our goal is to verify that the intuition from the synthetic setting carries over to a real dataset.

Semi-Synthetic Data Generation: We generated the dataset for this experiment as follows.

(a) Recovery of inclusion rule

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	1.0	1.0	0
$\alpha = 0.96$	1.0	1.0	0
$\alpha = 0.97$	1.0	1.0	1.0
$\alpha = 0.98$	1.0	1.0	1.0
$\alpha = 0.99$	1.0	1.0	1.0

(b) Avg. # of rules

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	23.67	15.75	5.0
$\alpha = 0.96$	35.58	33.33	4.0
$\alpha = 0.97$	39.83	31.92	4.0
$\alpha = 0.98$	44.17	47.17	23.83
$\alpha = 0.99$	31.42	31.25	27.67

(c) Avg. # of Perfect Rules

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	12.5	9.25	0.0
$\alpha = 0.96$	20.75	18.67	0.0
$\alpha = 0.97$	24.67	24.92	1.0
$\alpha = 0.98$	30.17	28.33	14.0
$\alpha = 0.99$	23.0	24.08	20.42

Table S2: Value of α parameter and λ_1 parameters, for a fixed beam search width ($B = 15$), along with (a) the proportion of runs (across all other hyperparameter settings) in which the synthetic region was correctly identified by the final rule set, (b) the number of rules in the final solution, and (c) the number of perfect rules, defined as those which exclude none of the samples but which exclude some number of reference points. Note that these results marginalize over λ_0, and (b-c) are averaged across all runs.

1. *Subsampling:* We randomly sample 5000 patients from the full cohort of 65k patients, due to computational constraints. In this subset, there were 185 binary features, and 5 continuous features.
2. *Synthetic Exclusion:* We remove all male patients with cardiac arrhythmia, which was around 5% of the total population.
3. *Pre-Processing:* Given the prevalence of very rare binary features, we removed all binary features with a prevalence of less than 1%, as well as all samples that had any of these features, resulting in the removal of 118 binary features and 850 samples. This was done both for computational reasons (to reduce the number of features) as well as to condition the problem such that it is more realistic for the support estimation to recover higher-order interactions.

4. *Final Dataframe*: The final dataset had 66 binary features and 5 continuous features, with the latter being converted into binary features via the use of deciles.

Hyperparameter Search: We then followed a similar approach to the synthetic experiment, using every combination of the following hyperparameters. For each combination, we ran the algorithm three times, inducing randomness over the data by taking a random 80% of the data with each iteration.

- $\alpha \in \{0.95, 0.96, 0.97, 0.98, 0.99\}$
- $\lambda_0 \in \{10^{-6}, 10^{-4}, 10^{-2}\}, \lambda_1 \in \{10^{-6}, 10^{-4}, 10^{-2}\}$

In this case, we fixed the width of the beam search at $B = 1000$ (which encourages a more thorough search during column generation, as discussed above), and also found that we needed to adjust the value of K, another hyperparameter from the column generation algorithm, to be roughly on the same order as B. The parameter K controls how many rules get added to the LP at each iteration. We also fixed the maximum number of iterations at 10. We recorded all the same outcomes as were used in the synthetic case.

Observations: We observed that a number of patterns from the synthetic case carried over to the semi-synthetic case.

- *Inclusion in LP (mostly) implies inclusion in final rules:* When the desired inclusion rule appears among the rules considered during column generation, it mostly appears in the final rounded rules, in 80% of runs. We conjecture that this is due to a large number of "perfect" rules existing in this dataset, which are also two-variable interactions, though many of these appear to be noise (see example inclusion rules below).
- *Increasing α leads to more consistent recovery in the LP* of the desired inclusion rule. However, as discussed, this does not always translate into the desired inclusion rule showing up in the final rounded rule set. See Table S3
- *Higher values of α are less sensitive to choice of λ*: In Tables (S4a-S4b) we demonstrate that, similar to the synthetic case, the number of rules and the number of "perfect" rules is highly sensitive to λ_1 when α is lower, but for $\alpha \geq 0.98$ it yields consistent results across different values of λ.

Example "Perfect" Rules: These rules exclude none of the samples in our data, while excluding reference points. While occasional rules appear to be based on reasonable exclusions (such as a lack of pregnant veterans, given that 80% of veterans are male in our data), most appear to be combinations of rare features (such as rare medications) that simply do not appear together in our data. These are three representative rules from one run (where $\alpha = 0.99, \lambda_0 = \lambda_1 = 1e-6$, resulting in 23 rules, of which 17 were "perfect"):

- not (Pregnant and Veteran)
- not (Complicated Hypertension and Previous Medication of Cephalexin)
- not (Previous Medication of Doxycycline and Norfloxacin)

Table S3: Values of α and the proportion of runs in which the desired inclusion rule was included in the LP during column generation, as well as included in the final rule set. Results are averaged over values of λ_0, λ_1, with the exception of $\lambda_0 = \lambda_1 = 1e-2$, because this did not run for $\alpha = 0.97$

	LP	Final Rule Set
$\alpha = 0.95$	0.50	0.50
$\alpha = 0.96$	0.75	0.71
$\alpha = 0.97$	1.00	0.88
$\alpha = 0.98$	1.00	0.62
$\alpha = 0.99$	1.00	0.62

(a) Recovery of inclusion rule

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	0.7	1.0	0.0
$\alpha = 0.96$	1.0	0.8	0.0
$\alpha = 0.97$	0.8	0.7	1.0
$\alpha = 0.98$	0.7	0.5	0.7
$\alpha = 0.99$	0.8	0.7	0.3

(b) Avg. # of rules

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	210.2	115.8	6.0
$\alpha = 0.96$	334.3	148.0	5.0
$\alpha = 0.97$	25.2	75.2	49.8
$\alpha = 0.98$	25.0	24.7	24.3
$\alpha = 0.99$	23.3	23.3	23.7

(c) Avg. # of Perfect Rules

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	200.2	105.8	0.0
$\alpha = 0.96$	326.0	140.0	0.0
$\alpha = 0.97$	19.5	69.0	42.2
$\alpha = 0.98$	21.3	21.0	20.7
$\alpha = 0.99$	19.0	18.7	19.7

Table S4: Value of α parameter and λ_1 parameters, along with (a) the proportion of runs (across all other hyperparameter settings) in which the synthetic region was correctly identified by the final rule set, (b) the number of rules in the final solution, and (c) the number of perfect rules, defined as those which exclude none of the samples but which exclude some number of reference points. Note that these results marginalize over $\lambda_0 \in \{1e-6, 1e-4\}$ because $\lambda_0 = \lambda_1 = 1e-2$ did not run for $\alpha = 0.97$, and (b-c) are averaged across all runs.

Table S5: **Rec**: Proportion of runs where synthetic exclusion was recovered. # **R**: Number of rules in final output. # **PR**: Number of "perfect" rules which exclude zero data points. **Length**: Average length of rules. Each entry is the average of three independent runs with different random seeds, and run with $B = 15$

α	λ_0	λ_1	Rec	# R	# PR	Length
0.95	0	1e-06	1.00	31.00	17.00	2.36
		1e-04	1.00	19.33	12.00	2.25
		1e-02	0.00	5.00	0.00	1.00
	1e-06	1e-06	1.00	30.67	17.00	2.37
		1e-04	1.00	19.33	12.00	2.25
		1e-02	0.00	5.00	0.00	1.00
	1e-04	1e-06	1.00	27.00	15.00	2.36
		1e-04	1.00	18.33	12.00	2.23
		1e-02	0.00	5.00	0.00	1.00
	1e-02	1e-06	1.00	6.00	1.00	1.17
		1e-04	1.00	6.00	1.00	1.17
		1e-02	0.00	5.00	0.00	1.00
0.96	0	1e-06	1.00	46.33	28.33	2.69
		1e-04	1.00	43.67	25.00	2.43
		1e-02	0.00	4.00	0.00	1.00
	1e-06	1e-06	1.00	45.33	27.67	2.70
		1e-04	1.00	43.67	25.67	2.41
		1e-02	0.00	4.00	0.00	1.00
	1e-04	1e-06	1.00	45.67	26.00	2.67
		1e-04	1.00	41.00	23.00	2.41
		1e-02	0.00	4.00	0.00	1.00
	1e-02	1e-06	1.00	5.00	1.00	1.20
		1e-04	1.00	5.00	1.00	1.20
		1e-02	0.00	4.00	0.00	1.00
0.97	0	1e-06	1.00	49.67	31.00	2.74
		1e-04	1.00	38.00	30.00	2.51
		1e-02	1.00	4.00	1.00	1.25
	1e-06	1e-06	1.00	49.67	31.00	2.73
		1e-04	1.00	38.00	30.00	2.51
		1e-02	1.00	4.00	1.00	1.25
	1e-04	1e-06	1.00	48.33	29.00	2.71
		1e-04	1.00	37.33	29.33	2.55
		1e-02	1.00	4.00	1.00	1.25
	1e-02	1e-06	1.00	11.67	7.67	2.27
		1e-04	1.00	14.33	10.33	2.43
		1e-02	1.00	4.00	1.00	1.25
0.98	0	1e-06	1.00	47.00	33.67	2.82
		1e-04	1.00	50.67	30.33	2.74
		1e-02	1.00	27.33	16.00	1.97
	1e-06	1e-06	1.00	46.67	33.33	2.81
		1e-04	1.00	50.67	30.33	2.74
		1e-02	1.00	27.00	15.67	1.97
	1e-04	1e-06	1.00	46.00	31.33	2.74
		1e-04	1.00	50.67	31.00	2.74
		1e-02	1.00	28.00	16.33	1.99
	1e-02	1e-06	1.00	37.00	22.33	2.29
		1e-04	1.00	36.67	21.67	2.26
		1e-02	1.00	13.00	8.00	1.95
0.99	0	1e-06	1.00	33.00	23.33	2.33
		1e-04	1.00	33.00	27.33	2.33
		1e-02	1.00	28.33	21.00	1.96
	1e-06	1e-06	1.00	33.00	21.67	2.36
		1e-04	1.00	34.33	24.67	2.30
		1e-02	1.00	28.33	21.00	1.96
	1e-04	1e-06	1.00	31.33	25.67	2.34
		1e-04	1.00	27.00	20.67	2.17
		1e-02	1.00	28.33	21.00	1.96
	1e-02	1e-06	1.00	28.33	21.33	2.08
		1e-04	1.00	30.67	23.67	2.11
		1e-02	1.00	25.67	18.67	1.96

B Application of OverRule to Policy Evaluation

In this section we give the detailed algorithm for applying OverRule to policy evaluation, as described in the main paper. In this context, we wish to evaluate not a specific treatment decision (e.g., the average treatment effect of giving a drug vs. withholding it), but rather a conditional *policy* representing a personalized treatment regime, which we will refer to as the *target* policy. This problem falls under the setting of off-policy policy evaluation when this target policy π differs from the policy which generated the data, which we observe in the observational data as $p(T = t \mid x)$.

Rationale for $\mathcal{B}^\epsilon(\pi)$: In the main paper, we drew a connection between the set $\mathcal{B}^\epsilon$ and the following set, a function of the target policy π, $\mathcal{B}^\epsilon(\pi) := \{x \in \mathcal{X}; \forall t : \pi(t \mid x) > 0 : p(T = t \mid x) > \epsilon\}$. In this section, we recall the theoretical rationale for why we are restricted to this set, if we wish to evaluate the policy π given samples generated according to $p(T = t \mid x)$.

Following similar notation to Kallus and Zhou (2018), we will let $X \in \mathcal{X}$ correspond to covariates, $Y \in \mathcal{Y}$ to an outcome of interest, $T \in \mathcal{T}$ to a treatment decision. We write $\pi(t|x_i)$ as the probability of each treatment under the policy, which may be stochastic. We write $Y(t)$ to represent the potential outcome under treatment t. In this setting, we wish to evaluate the expected value of Y under the target policy, which we denote as $\mathbb{E}[Y(\pi)]$.

Proposition S1 (Informal)**.** *The expectation $\mathbb{E}[Y(\pi)]$ is only defined w.r.t. the observed distribution $p(X, T, Y)$ for the subset $B \in \mathcal{X}$ such that $\forall x \in B$, $\pi(T = t \mid X = x) > 0 \implies p(T = t \mid X = x) > 0$*

Proof. Under the assumption that ignorability (Pearl, 2009) holds, we can write out our desired quantity as follows in terms of observed distribution $p(X, T, Y)$. For brevity, let $p(t \mid x) = p(T = t \mid X = x), p(x) = p(X = x)$, et cetera.

$$\mathbb{E}[Y(\pi)] \tag{S1}$$

$$= \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x)\pi(t \mid x) \cdot p(Y(t) = y \mid x, t)dxdtdy$$

$$= \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x)\frac{\pi(t \mid x)}{p(t \mid x)} \cdot p(Y(t) = y \mid x, t)p(t \mid x)dxdtdy \tag{S2}$$

$$= \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x)p(t \mid x) \cdot p(Y = y \mid x, t)\frac{\pi(t \mid x)}{p(t \mid x)}dxdtdy \tag{S3}$$

$$= \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x, t, y) \cdot \frac{\pi(t \mid x)}{p(t \mid x)}dxdtdy \tag{S4}$$

Where in Equation (S2) we multiply by one, in Equation (S3) we use the assumption of ignorability to write $p(Y(t) = y \mid X = x, T = t) = p(Y = y \mid X = x, T = t)$ and rearrange terms, and in Equation (S4) we collect the terms which represent the observed distribution. For our purposes, it is sufficient to look at the integral in Equation (S4) to see that it requires the condition that for all $(x, t) \in \mathcal{X} \times \mathcal{T}$, the relationship $\pi(T = t \mid X = x) > 0 \implies p(T = t \mid X = x) > 0$ must hold. □

The condition given in Proposition S1 is sometimes referred to as the condition of *coverage* (see Sutton and Barto, 2017, Section 5.5) in off-policy evaluation. Rewriting Equation (S4) as an expectation over the observed distribution, we can see that this leads naturally to the importance sampling (Kahn, 1955) estimator

$$\mathbb{E}\left[Y\frac{\pi(T = t \mid X = x)}{p(T = t \mid X = x)}\right] \approx \frac{1}{n}\sum_{i=1}^{n} y_i \frac{\pi(t_i \mid x_i)}{p(t_i \mid x_i)}, \tag{S5}$$

which approximates our desired quantity. If $\epsilon > p(t|x) > 0$ for some small value of ϵ, then the variance of the importance sampling estimator increases dramatically. This motivates our notion of "strict" coverage, that for each value of $x \in \mathcal{B}^\epsilon(\pi)$, we require that for all actions t such that $\pi(t|x) > 0$, the condition $p(t|x) > \epsilon$ must hold.

Note that this differs conceptually from the binary treatment case in an important respect: Since we are not seeking to contrast all treatments, we do not require that $\mu(t|x) > \epsilon, \forall t \in \mathcal{T}$, but rather just for those treatments which have positive probability of being taken under the target policy.

Algorithmic Details As described in the main paper, applying OverRule to the policy evaluation setting only requires a single change to the procedure, which is that the set $\hat{B}^\epsilon(\pi)$ is used in place of the set $\hat{B}^\epsilon$ in Equation (9) in Section 4.2. Nonetheless, we provide an explicit self-contained sketch of the procedure here to avoid any confusion:

1. Given a dataset, find an α-MV set $\mathcal{S}^\alpha$ using the approach given in the main paper.

2. Using this set, learn the conditional probabilities of each possible treatment $t \in \mathcal{T}$, resulting in estimated propensities $\hat{p}(T = t \mid X = x)$

3. For each data point in the support set $\mathcal{S}^\alpha$, assign the label

$$\hat{b}_i(\pi) = \prod_{t \in \pi(x_i)} \mathbb{1}[\hat{p}(T = t \mid X = x_i) \geq \epsilon],$$

where $\pi(x_i) \coloneqq \{t : \pi(t|x_i) > 0\}$. The set $\hat{B}^\epsilon(\pi)$ is the collection of data points such that $\hat{b}_i(\pi) = 1$. Note that we know the target policy π that we are evaluating, so we can evaluate $\pi(t|x_i)$ for each data point.

4. Solve the following Neyman-Pearson-like classification problem, using the techniques discussed in the main paper. Note that this is identical to solving Equation (9) in Section 4.2, with the substitution of $\hat{B}^\epsilon(\pi)$ for $\hat{B}^\epsilon$:

$$\hat{\mathcal{B}}(\pi) := \arg\min_{C} \frac{1}{|\hat{\mathcal{S}} \setminus \hat{B}|} \sum_{i \in \hat{\mathcal{S}} \setminus \hat{B}^\epsilon(\pi)} \mathbb{1}[x_i \in \mathcal{C}] + R(\mathcal{C})$$
$$\text{s.t.} \sum_{i \in \hat{\mathcal{S}} \cap \hat{B}^\epsilon(\pi)} \mathbb{1}[x_i \in \mathcal{C}] \geq \beta |\hat{\mathcal{S}} \cap \hat{B}^\epsilon(\pi)| \ .$$

C Additional Experimental Results

As a general note across all experiments: When estimating support in OverRule, we use $m_R = c \cdot m \cdot d$ uniform reference samples where $c > 0$ is some constant, m is the number of data samples and d their dimension. Continuous features were binarized by deciles unless otherwise specified. Finally, for propensity-based base estimators, we use the standard threshold $\epsilon = 0.1$ (Crump et al., 2009) throughout.

C.1 Iris

For the results given in the paper, we fit OverRule using a k-NN base estimator ($k = 8$) and DNF Boolean rules for both support and overlap rules, with $\alpha = 0.9$ and regularization $\lambda_0 = 2 \cdot 10^{-2}, \lambda_1 = 0$ for support rules, a cutoff of $\epsilon = 0.1$, and $\beta = 0.9, \lambda_0 = 10^{-2}, \lambda_1 = 0$ for overlap rules.

C.2 Jobs

For the results given in the main paper, we use the following hyperparameters:

1. **Support Rules**: CNF formulation, along with hyperparamters $\alpha = 0.98, \lambda_0 = 10^{-2}, \lambda_1 = 10^{-3}$.
2. **Base Estimators**: For CBB we used $\alpha = 0.1$, for the logistic regression propensity estimator we used $C = 1$ in `LogisticRegression` in scikit-learn, and other hyperparameters were chosen based on cross-validation: For k-NN, we selected $k \in \{2, 4, \ldots, 20\}$ based on held-out accuracy in predicting group membership and used $1/k$ as threshold. For OSVM, we use a Gaussian RBF-kernel with bandwidth $\gamma \in [10^{-2}, 10^2]$, selected based on the held-out likelihood of kernel density estimation.
3. **Overlap Rules**: We use a DNF formulation with $\beta = 0.9$ and select $\lambda_0 \in [10^{-4}, 10^{-1}]$ and $\lambda_1 \in [10^{-4}, 10^{-2}]$. Within each class of base estimators, we choose these parameters based on average *training* performance over 5-fold CV, choosing the setting in each class that achieves a balanced accuracy (with respect to the base-estimator overlap labels) within 1% of the best performing model in the class, while minimizing the number of rules.

Note that the reported results are using the held-out portions of each 5-fold CV run, and using the ground-truth overlap labels, which are at no point used during the hyperparameter tuning process. This reflects a real-world scenario where ground-truth is unknown and only the base-estimator derived labels are given. The reported rules in the figure were selected from one of the five cross-validation runs for the same hyperparameter setting chosen using the above procedure. In Figure S2 we see the correlation between held-out balanced accuracy for the rule set w.r.t. the experimental label, and the balanced accuracy for the rule set in approximating the base estimator. Note that AUC is equal to balanced accuracy for binary predictions.

C.3 Opioids

For the results in the main paper, we fit an OverRule model (OR) to a random forest base estimator with $\beta = 0.8$ for $\mathcal{B}$ and $\alpha = 0.9$ for $\mathcal{S}$ picked a priori. The hyperparameter λ_0 was set to $\lambda_0 =$1e-3 for $\mathcal{B}$, and $\lambda_0 =$1e-5 for $\mathcal{S}$, and $\lambda_1 = 0$ for both.

For a full table of covariate statistics for the Opioids dataset, see Table S6. For a illustration of the rules learned by OverRule to describe the complement of the overlap set, see Figure S3.

Supplemental Rules: We learned an additional set of rules, motivated by our experiments in Section 5.3, where we noted that the support rules did not capture certain combinations of surgery types or conditions that should be rare or non-existent. This motivated the empirical investigation in Section A.2, and this vignette represents the result of re-running our procedure with this goal in mind.

For support rules, we followed the recommendations laid out in Section A.2, choosing to use a CNF formulation with $\alpha = 0.98, \lambda_0 = 0, \lambda_1 = 0.01$. Continuous features were binarized using deciles. For our base estimator, we used a random forest classifier with 100 trees and 20 minimum samples per leaf, and we used $\epsilon = 0.1$ as our cutoff. For the overlap rules, we searched over the following grid of hyperparameters, with the goal of maximizing balanced accuracy with respect to the overlap labels on a validation set:

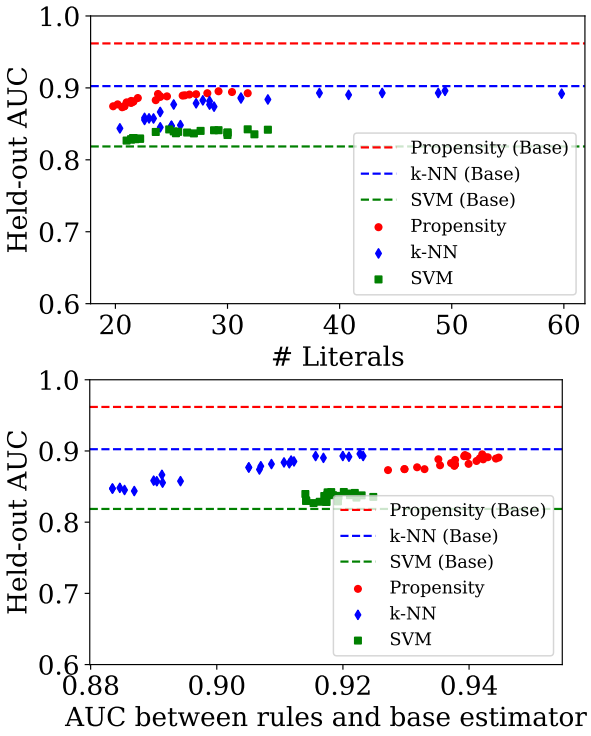

Figure S2: Results from the Jobs datasets for OverRule approximations of different base estimators, sweeping λ_0, λ_1. AUC (i.e., balanced accuracy) is measured with respect to the experimental indicator. The dotted line 'Propensity (base)' refers to the logistic regression base estimator, 'k-NN (base)' refers to the k-NN base estimator, and 'SVM (base)' refers to the one-class SVM. The colored points refer to performance of OverRule using the respective base estimator, for different values of λ_0, λ_1

$\beta \in \{0.8, 0.9, 0.95\}$ and then a set where $\lambda_0 = 0$ and $\lambda_1 \in \{10^{-3}, 2 \cdot 10^{-3}, 10^{-2}\}$, and a set where $\lambda_1 = 0$ and $\lambda_0 \in \{10^{-3}, 2 \cdot 10^{-3}, 10^{-2}\}$. The selected hyperparameters were $\beta = 0.95, \lambda_0 = 0, \lambda_1 = 10^{-3}$. The support rules cover 98.5% of the test samples, and the overlap rules achieved a balanced accuracy of 0.96 on a held-out test set (with respect to the overlap labels) and covered 36% of the test samples. The chosen ruleset is given in given in Figures S4-S5.

We note that the resulting support rules, in line with the findings in Section A.2, include a large number of rules that exclude zero training data points, by identifying rare interactions of features. For instance, the rules identify that there are *no men in our dataset who have maternity surgery*, an intuitive exclusion.

We shared this rule set with one of the participants of the original user study, who made the following observations: First, the support rules in Figure S4 generally made sense as excluding combinations that are intuitively absent from the data (e.g., men w/maternity surgery) or that are just combinations of features that are themselves rare. Regarding the overlap rules in Figure S5, they observed that B.1 and B.2 were consistent with clinical intuition, where B.2 likely serves to exclude C-section patients with epidurals. B.3 and B.4 were intuitive with the exception of the negations, e.g., it is unclear what the role of abdominal pain is in B.3, although it could be correlated with generalized pain syndromes. B.5-B.7 correspond to individuals with lower back pain (Lumbago) and neck pain (Cervicalgia) which are intuitive indicates for higher doses of opioids. B.8 corresponds to plastic surgery, and the broad category of respiratory surgery in B.9 could correspond to thoracic surgery, one of the main surgical categories associated with opioid misuse. B.10-B.12 relate to back pain, which is associated with higher opioid dosages.

C.4 Observational Study: Policy Evaluation of Antibiotic Prescription Guidelines

Antibiotic resistance is a growing problem in the treatment of urinary tract infections (UTI) (Sanchez et al., 2016), a common infection for which more than 1.6 million prescriptions are given annually in the United States (Shapiro et al., 2013). With this in mind, we are interested in the following clinical problem: When a patient presents with a UTI, the physician needs to choose between a range of antibiotics, with the dual goals of (a) treating the infection, and (b) minimizing the use of broad-spectrum antibiotics, which are more likely to select for drug-resistant strains of bacteria.

In this context, we might be interested in evaluating a range of potential treatment policies. For our purposes, we will use a pre-defined policy: The clinical guidelines published by the Infectious Disease Society of America (IDSA) for treatment of uncomplicated UTIs in female patients (Gupta et al., 2011). Using the policy evaluation formulation of $\mathcal{B}^{\epsilon}(\pi)$, we will apply OverRule to a conservative interpretation of the IDSA guidelines, using data curated from the Electronic Medical Record (EMR) of two academic medical centers.

The official guidelines discuss the importance of patient and population level risk factors in predicting resistance, and include some factors that we do not observe in our data (such as drug allergies). In order to characterize the guideline explicitly as a policy that we can evaluate in our dataset, we used the following interpretation:

- Choose the first-line agent, either Nitrofurantoin (NIT) or Trimethoprim/Sulfamethoxazole (SXT), to which the patient did not have previous antibiotic exposure or resistance in the prior 90 days. Additionally, if local rates of resistance to SXT are $\geq$ 20% in the prior 30-90 days, then avoid

Support rules $\hat{\mathcal{S}}$

Rule S.1:

	History:
	¬ Injury of face and neck
and	¬ Unspecified septicemia
and	¬ Other injury of chest wall
and	¬ Acute respiratory failure
and	¬ Altered mental status
and	**Surgical procedure:**
	¬ Endocrine system
and	¬ Mediastinum (thoracic cavity)
and	¬ Auditory system

Propensity overlap complement rules $\widehat{\mathcal{B}^c}$

Rule B.1:

	Surgical procedure:
	¬ Respiratory
and	¬ Nervous
and	¬ Musculoskeletal
and	¬ Cardiovascular
and	**History:**
	¬ Tobacco use disorder
and	¬ Thoracic or lumbosacral neuritis or radiculitis: unspecified
and	¬ Lumbosacral spondylosis without myelopathy
and	¬ Degeneration of cervical intervertebral disc
and	¬ Degeneration of lumbar or lumbosacral intervertebral disc

or Rule B.2:

	Surgical procedure:
	Maternity
and	**History:**
and	¬ Degeneration of lumbar or lumbosacral intervertebral disc

$\hat{\mathcal{O}} = \text{S.1} \wedge \neg\,(\text{B.1} \vee \text{B.2})$

Figure S3: OverRule description of the *complement* of the overlap between post-surgical patients with higher and lower opioid prescriptions. If the support rule (left) applies and *neither* propensity overlap rule (right) applies, a patient is consider to be in the overlap set. ¬ indicates a negation. The rules cover 36% of patients with balanced accuracy 0.92 w.r.t. the base estimator (random forest). Procedures are not mutually exclusive.

prescription of SXT.

- If neither of the first-line agents are indicated, then prescribe Ciprofloxacin (CIP), a second-line agent.

Experimental details From our data set, we selected all patients from 2007–2017 which had a UTI, and were prescribed one of the four most common antibiotics: NIT, SXT, CIP, or Levofloxacin (LVX). Features include demographics (race, gender, age, and veteran status), comorbidities observed in the past 90 days, information about previous infections (organism, antibiotics given, and resistance profile), hospital ward (inpatient, outpatient, ER, and ICU), and indicators for pregnancy and nursing home residence in the past 90 days. The local rates of resistance (for each hospital ward) are given over the past 30–90 days, and used at the patient level as a feature, as well as an input to the decision of the guidelines.

We preprocess our data first, removing any binary feature with a prevalence of less than 0.1%, and any associated subject: This results in the removal of 48 binary features with less than 0.1% prevalence and 888 corresponding subjects. This leaves a total of 156 (150 binary, 6 continuous) features and 64593 subjects. Detail on all remaining features are given in Table S7. For the purposes of running our algorithm, we convert all continuous variables into binary variables by using indicator functions for deciles.

We then characterize the support set $\mathcal{S}^\alpha$ as described in the main paper, using a DNF formulation, along with $\alpha = 0.95, \lambda_0 = 0.01, \lambda_1 = 0$. Using the data points which fall into the support set, we then estimate the propensity $p(t|x)$ of prescribing each of the four drugs using a random forest classifier, with hyperparameter selection done using 5-fold cross-validation on 80% of the remaining cohort used as a training set, over the following parameter grid: Number of estimators $\in [100, 500]$, Minimum samples per leaf (as fraction of total) $\in [0.005, 0.01, 0.02]$. The resulting calibration curves for each antibiotic are given in Figure S6, using the remaining held-out 20% of the data. Using these propensity scores, we apply the procedure described in Section B to estimate the region of strict coverage, $\hat{\mathcal{B}}^\epsilon(\pi)$ using Boolean rules, and the resulting rules are given in Figure S7. For this stage, we used a DNF formulation and hyperparameters of $\beta = 0.9, \lambda_0 = 0.03, \lambda_1 = 0$.

Clinical Validity / Interpretation Towards understanding the clinical validity of these rules, we interviewed a clinician who specialises in infectious diseases. First, we asked them, based on the available features, which they would expect to differentiate between subjects for whom the policy is or is not followed. They noted that the guidelines are designed for uncompli-

Support rules $\widehat{S}$

NONE OF:

Proc: Auditory

Hist: Unspecified septicemia

Hist: Acute respiratory failure

Male
and **Proc:** Maternity

Male
and **Proc:** Female Genital

Hist: Other screening mammogram
and **Proc:** Male Genital

Proc: Musculoskeletal
and **Proc:** Male Genital

Proc: Respiratory
and **Proc:** Female Genital

Hist: Injury of face and neck
and **Proc:** Male Genital

Hist: ADD (w/hyperactivity)
and **(Hist:** Rheumatoid arthritis
OR **Hist:** Other symptoms referable to back
OR **Hist:** Myalgia and myositis: unspecified**)**

Hist: ADD (without hyperactivity)
and **(Hist:** Rheumatoid arthritis
OR **Hist:** Other symptoms referable to back
OR **Proc:** Male Genital)

Hist: Major depressive affective disorder
and **(Hist:** Other symptoms referable to back
OR **Proc:** Male Genital)

Hist: Hypopotassemia
and **Hist:** Hypersomnia with sleep apnea

Hist: Injury of face and neck
and **Proc:** Fitting and adjust. of vascular catheter

Figure S4: Support Rules using CNF formulation for the Opioids task. **Proc** indicates a procedure, and **Hist** indicates a history of a condition. A sample is considered in the support set if NONE of the above rules apply. Note that rules are negated for simplicity of presentation, as "AND NOT (X AND Y)" is equivalent to "AND (NOT X OR NOT Y)", and in some cases several rules are combined for simplicity of presentation (e.g., those related to Attention Deficit Disorder). Dark green rules are highlighted to indicate that they cover <4 training samples (and in many cases zero training samples) in line with our findings in Section A.2 for this setting of hyperparameters.

cated cases: In particular, patients who have a Foley catheter (a catheter used to drain urine from the bladder) are not covered under these guidelines, because infections in these patients tend to be more complex (e.g., the infection could have been introduced by the catheter itself). The use of the Foley catheter is common during intensive care (e.g., in the ICU), so complex hospitalized patients are less likely to be treated according to the guidelines.

With that in mind, they reviewed the available features and noted the following: (i) While UTIs are common for women, they are rare for men; Men with UTIs tend to be more complicated cases, because it is indicative of deeper abnormalities. Similarly, pregnant women are excluded from the guidelines. (ii) Of the comorbidities given, none of them should directly disqualify patients from the guidelines, except potentially for complicated diabetes. (iii) Prior organisms / resistance / prescriptions should not directly disqualify patients from the guidelines, though they will influence the type of antibiotic given. In particular, if a patient has had previous resistance to an antibiotic, they are unlikely to be prescribed it again. (iv) The previous procedures given (with the exception of surgery) are associated with ICU patients. For instance, mechanical ventilation and parenteral nutrition are exclusive to the ICU, and those patients likely have a Foley catheter as well. Surgery is too broad of a category to draw any conclusions. (v) In terms of locations besides the ICU, patients who are admitted to the hospital and who are on intravenous (IV) antibiotics already will be treated differently. The guidelines are focused on oral antibiotics, whereas if an IV already exists, additional IV antibiotics are likely to be given instead.

Having discussed these points first, we then showed them the rules learned by the OverRule algorithm, and asked for their interpretation, as well as for any critiques of the rules based on their clinical knowledge. Their reaction to each of the rules was as follows:

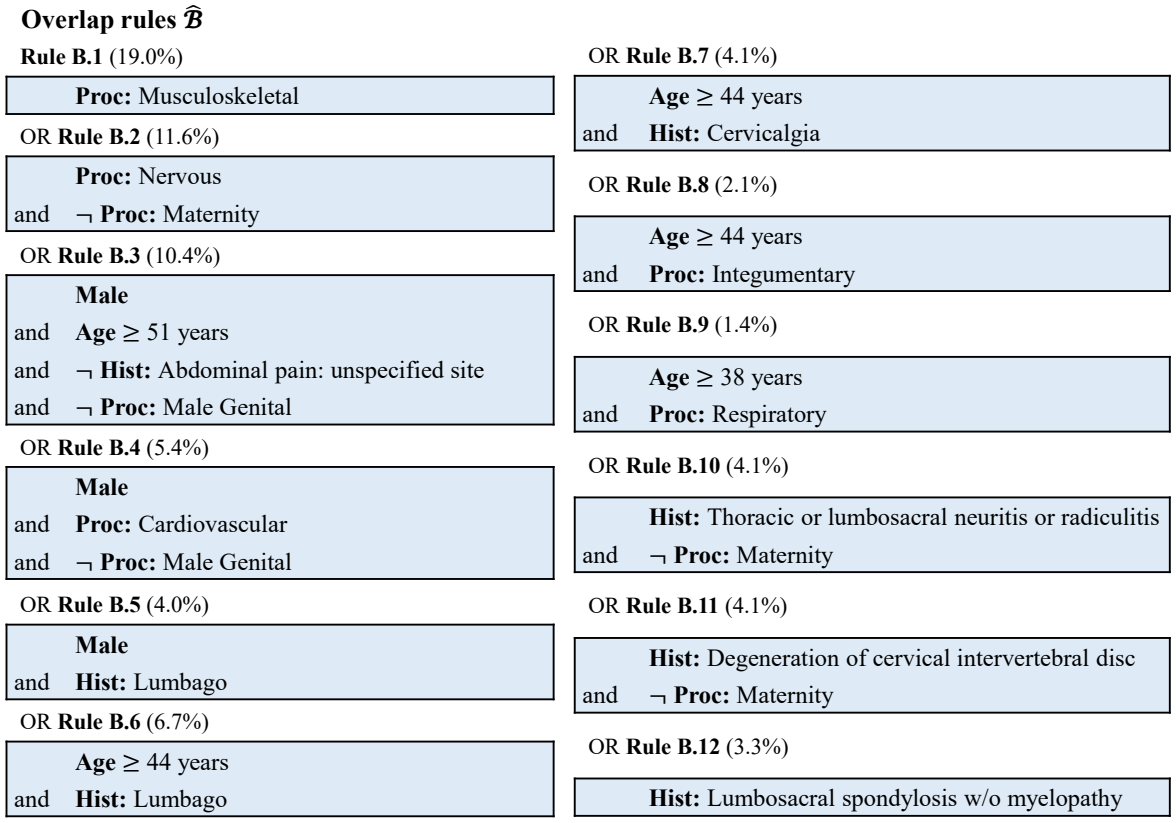

Figure S5: Overlap rules, where the percentage next to each rule indicates the percentage of the dataset that is covered by that rule. Collectively, these rules cover 36% of the held-out datapoints.

- **Rule B.1**: This appears to correspond to a relatively straightforward young inpatient female (given that Rule B.2 covers all outpatient females). In particular, it rules out ICU patients directly, as well as those with recent mechanical ventilation, which would indicate a recent ICU stay. It also rules out patients with current bloodstream infections, and those who had previously been tested for (and found to be) resistance to Streptomycin (synergistic): This is only tested for in the context of bloodstream infections by enterococcus, and would be an indicator of previous bloodstream infections. Imipenem is an IV antibiotic only given in inpatient settings, and posaconazole is an antifungal used in bone marrow transplant patients. Patients who are both young and in a nursing home tend to be more complex, e.g., they may be paralysed or otherwise unable to perform activities independently. Finally, the excluded comorbidities are less intuitive, because some of them (e.g., congestive heart failure) manifest with a range of severity: For patients with controlled congestive heart failure, this is not a contraindication for following the guidelines, but if they are fully decompensated, then they would likely be on a Foley catheter.

- **Rule B.2**: This concisely describes the most common manifestation of UTI and the set of patients who are most likely to be treated according to the guidelines[3].

- **Rule B.3**: The conjecture is that this represents patients who have had an uncomplicated UTI in the past, since patients are usually tested for the antibiotics under consideration by a physician, and since nitrofurantoin is one of the first-line treatments for uncomplicated UTIs.

From a quantitative perspective, we compared the learned region with an explicitly constructed cohort of patients whose inclusion criteria were explicitly designed to make them eligible for application of the IDSA guidelines. In particular, we defined this cohort as including non-pregnant women between the ages of 18 to 55 years of age with no record of genitourinary surgery or instrumentation, immunosuppression, indwelling catheters, or neurologic dysfunction in the

[3]Note that outpatient and "not inpatient" can appear in the same rule without being redundant, because multiple specimens collected on the same day for the same patient are collapsed into a single subject.

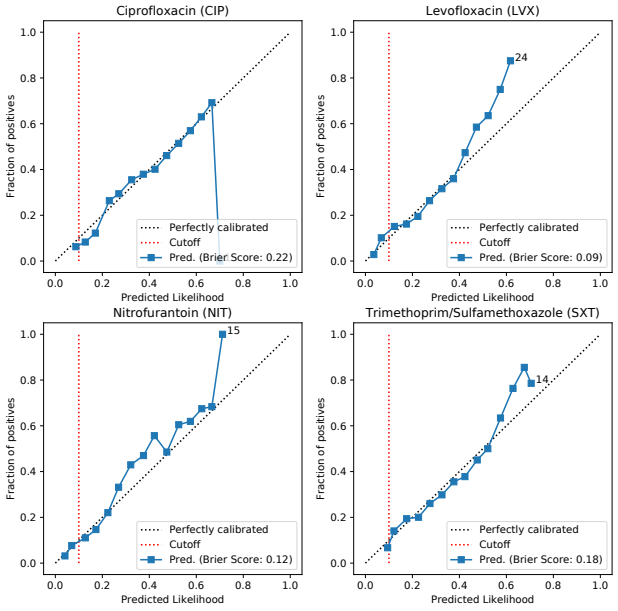

Figure S6: Calibration curves for each antibiotic, using 20 evenly spaced bins in the range $[0, 1]$. Numbers indicate the number of samples, and are given when when the number of samples in a bin is less than 0.5% of the total. The cutoff is a reminder that $\epsilon = 0.1$ in this experiment: For any subject with covariates x, the propensity must be above this cutoff for every treatment under the target policy (i.e., for all t such that $\pi(t|x) > 0$) for them to be included in the coverage region.

preceding 90 days. There were 14k of these patients, 21% of the total.

In relationship to this conservative subset, the learned region (covering 42k patients, 64% of total) covers 96% of the explicitly constructed cohort, while also demonstrating that a broader set of patients are treated according to these guidelines in practice.

D Theoretical Results on Regularized Minimum-Volume Boolean Rules

D.1 Bounds on minimum volume

In this subsection, we derive lower bounds on the volume of optimal DNF Boolean rules in problem (5).

First we obtain an expression for the normalized volume of a clause in a DNF (we use the terms clause and conjunction interchangeably in the case of a DNF). We express the domain $\mathcal{X}$ as the Cartesian product $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. A DNF rule with K clauses a_k is written as

$$r(x) = \bigvee_{k=1}^{K} a_k(x) = \bigvee_{k=1}^{K} \bigwedge_{j \in \mathcal{J}_k} (x_j \in \mathcal{S}_{jk}), \tag{S6}$$

where $\mathcal{J}_k$ is the set of covariates participating in clause k, and each $x_j \in \mathcal{S}_{jk} \subseteq \mathcal{X}_j$ is a subset membership condition on an individual covariate. Examples of such conditions are (Age ≥ 30) for a continuous-valued covariate and (Sex $=$ Female) for a discrete-valued one. For $j \notin \mathcal{J}_k$, it is understood that $x_j \in \mathcal{X}_j$, i.e. there is no restriction on x_j. The volume of clause a_k is then given by the product

$$V(a_k) = \prod_{j \in \mathcal{J}_k} |\mathcal{S}_{jk}| \prod_{j \notin \mathcal{J}_k} |\mathcal{X}_j|,$$

where $|\mathcal{S}_{jk}|$ is the length of subset $\mathcal{S}_{jk}$ for a continuous covariate j or the cardinality of $\mathcal{S}_{jk}$ for a discrete covariate, and similarly for $|\mathcal{X}_j|$. Likewise, the volume of $\mathcal{X}$ is $\prod_{j=1}^{d} |\mathcal{X}_j|$, and the normalized volume of a_k is therefore

$$\bar{V}(a_k) = \prod_{j \in \mathcal{J}_k} f_{jk}, \quad f_{jk} = \frac{|\mathcal{S}_{jk}|}{|\mathcal{X}_j|} \in [0, 1]. \tag{S7}$$

We define $p_k = |\mathcal{J}_k|$ to be the *degree* of conjunction k.

Proposition S2. *Assume that the regularization $R(r)$ follows (6). Then in any optimal solution to (5), all clauses a_k of degree p_k have normalized volume satisfying $\bar{V}(a_k)^{(p_k-1)/p_k} - \bar{V}(a_k) \geq \lambda_1$.*

Proof. Suppose that rule r with corresponding set $\mathcal{C}$ is an optimal solution to (5). Recalling the expansion in (S6), we consider modifications to r in which one condition $(x_j \in \mathcal{S}_{jk})$ is removed from a clause a_k. The modified rule satisfies the mass constraint $P(\mathcal{C}) \geq \alpha$ because it covers at least those points covered by r. From (S7), the increase in volume is at most $\bar{V}(a_k)((1/f_{jk}) - 1)$, with equality if none of the additional volume is already covered by another clause in r, while the complexity penalty decreases by λ_1. The change in objective value is thus bounded from above by

$$\bar{V}(a_k)\left(\frac{1}{f_{jk}} - 1\right) - \lambda_1.$$

This upper bound must be non-negative as otherwise r is not optimal. In particular, for $f_{jk} = \max_{j' \in \mathcal{J}_k} f_{j'k}$ and all k we have

$$\bar{V}(a_k)\left(\frac{1}{\max_{j \in \mathcal{J}_k} f_{jk}} - 1\right) \geq \lambda_1.$$

Since (S7) implies that $\max_{j \in \mathcal{J}_k} f_{jk} \geq \bar{V}(a_k)^{1/p_k}$, the desired result follows. □

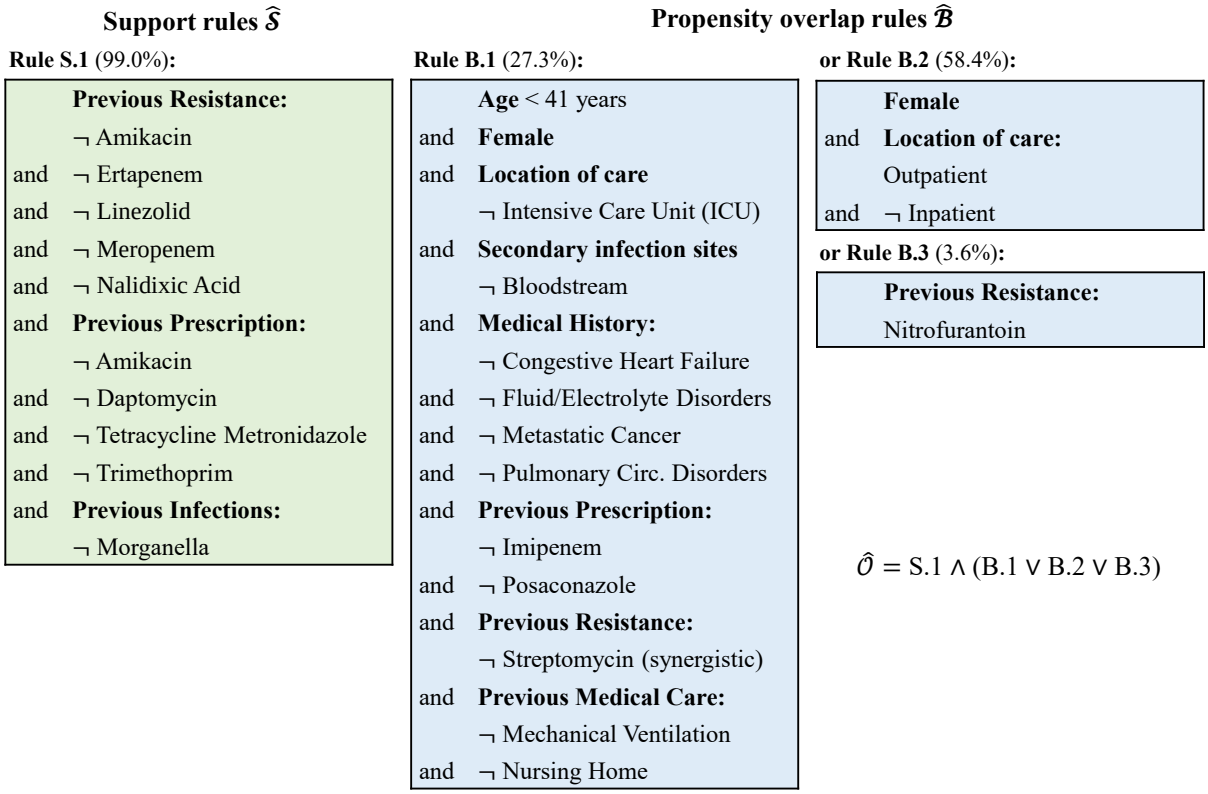

Figure S7: OverRule description of the coverage region for policy evaluation of the clinical guidelines. Beside each rule we give the percentage of subjects that are covered by the rule in the test set. Overall, the rules for $\hat{B}$ cover 65.4% of the data points in the support region (compared to the 71% of points labelled by our base estimator), and they have an balanced accuracy of 0.96 versus the base estimator.

For $p > 1$, the function $\bar{V}^{(p-1)/p} - \bar{V}$ is positive and concave on $(0, 1)$ with roots at 0 and 1. For $\lambda_1 > 0$, the equation $\bar{V}^{(p-1)/p} - \bar{V} = \lambda_1$ therefore has either two roots, $0 < \bar{V}_L < \bar{V}_U < 1$, which define an interval where the inequality $\bar{V}^{(p-1)/p} - \bar{V} \geq \lambda_1$ is satisfied, or no roots if λ_1 is too large. We are interested primarily in the root $\bar{V}_L$ as a lower bound on volume. While $\bar{V}_L$ is not available in closed form for $p > 2$, the following corollary gives a simple expression that is a lower bound on $\bar{V}_L$.

Corollary S1. *Under the assumption in Proposition S2, in any optimal solution to* (5)*, all clauses a_k of degree $p_k > 1$ have normalized volumes of at least $\lambda_1^{p_k/(p_k-1)}$.*

Proof. Proposition S2 implies $\bar{V}(a_k)^{(p_k-1)/p_k} \geq \lambda_1$ after dropping $-\bar{V}(a_k)$ from the left-hand side. □

Lastly, since the volume of a DNF rule is at least that of any of its clauses, we have the following.

Corollary S2. *Under the assumption in Proposition S2, any optimal solution to* (5) *has normalized volume of at least $\lambda_1^{p_{\max}/(p_{\max}-1)}$, where $p_{\max} = \max_k p_k$ is the largest degree of its clauses.*

D.2 Bounds on the number of candidate DNF rules

The results in the previous subsection are necessary conditions of optimality for problem (5). The implication is that in searching for optimal solutions to (5), we may restrict the class $\mathcal{C}$ of DNF rules considered to those satisfying these necessary conditions. In this subsection, we develop the consequences of this restriction, culminating in a bound on $|\mathcal{C}|$, the number of candidate DNF rules (Lemma S5).

For simplicity, we assume in the following that all variables X_j are binary-valued. An extension to non-binary categorical variables and continuous variables (discretized using interval conditions $l_j \leq x_j \leq u_j$) is likely possible with the additional complications of accounting for the cardinalities of categorical variables and bounding the fractions f_{jk} associated with continuous variables.

First, the simplified lower bound on volume in Corollary S1 implies an upper bound on conjunction degree.

Lemma S1. *Assume that the regularization $R(r)$ follows* (6) *and that all variables are binary. Then in any optimal solution to* (5)*, the maximum degree of a conjunction is $p_{\max} := 1 + \lfloor \log_2(1/\lambda_1) \rfloor$.*

Proof. The normalized volume of a conjunction of degree p_k is 2^{-p_k}. Corollary S1 then requires

$$2^{-p_k} \geq \lambda_1^{p_k/(p_k-1)}.$$

Taking logarithms and rearranging, we obtain

$$-1 \geq \frac{1}{p_k - 1} \log_2 \lambda_1,$$
$$p_k \leq 1 + \log_2(1/\lambda_1).$$

The right-hand side can be rounded down since p_k is integer. □

Given Lemma S1, we may enumerate DNF rules satisfying the lemma according to the numbers of clauses of degree $p = 1, \ldots, p_{\max}$ that they possess. Denote by K_p the number of clauses of degree p and call $\mathbf{K} = (K_1, \ldots, K_{p_{\max}})$ the *signature* of a DNF rule. The signatures of optimal DNF rules obey the following constraint.

Lemma S2. *Under the assumptions of Lemma S1, the signature $\mathbf{K} = (K_1, \ldots, K_{p_{\max}})$ of an optimal solution to* (5) *must satisfy*

$$\sum_{p=1}^{p_{\max}} K_p(\lambda_0 + p\lambda_1) < 1. \tag{S8}$$

Proof. From (6), the complexity penalty of a solution with K_p clauses of degree p, $p = 1, \ldots, p_{\max}$ is given by the left-hand side of (S8). For a solution to be optimal, it must have lower cost than the trivial "all true" rule, which has a normalized volume of 1 and complexity penalty of 0. In particular, the complexity penalty must be less than 1. □

Let Δ denote the set of signatures that satisfy (S8), and for $\mathbf{K} \in \Delta$, let $\mathcal{C}(\mathbf{K})$ be the set of DNF rules with signature $\mathbf{K}$. The number of DNF rules satisfying the necessary conditions of optimality in Lemmas S1 and S2 can be bounded as follows:

$$|\mathcal{C}| = \sum_{\mathbf{K} \in \Delta} |\mathcal{C}(\mathbf{K})| \leq |\Delta| \max_{\mathbf{K} \in \Delta} |\mathcal{C}(\mathbf{K})|. \tag{S9}$$

The next two lemmas provide upper bounds on the two right-hand side factors in (S9).

Lemma S3. *The number of signatures satisfying* (S8) *is bounded as*

$$|\Delta| \leq 2 \left(\frac{1}{\lambda_1} \right)^{p_{\max}}.$$

Proof. For simplicity, we consider a superset $\Delta_0 \supseteq \Delta$ obtained by dropping λ_0 from (S8), i.e.

$$\sum_{p=1}^{p_{\max}} p\lambda_1 K_p \leq 1. \tag{S10}$$

Condition (S10) together with the implicit non-negativity constraints $K_p \geq 0$, $p = 1, \ldots, p_{\max}$ define a simplex in $p_{\max}$ dimensions. Bounding the number of signatures in Δ_0 is thus equivalent to bounding the number of non-negative integer points in this simplex. This problem has been studied extensively by mathematicians. Applying e.g. (Yau and Zhang, 2006, eq. (1.5)), we have

$$\begin{aligned}
|\Delta_0| &\leq \frac{1}{p_{\max}!} \prod_{p=1}^{p_{\max}} \frac{1}{p\lambda_1} \left(1 + \sum_{p=1}^{p_{\max}} p\lambda_1 \right)^{p_{\max}} \\
&= \frac{1}{\left(p_{\max}!\right)^2} \left(\frac{1}{\lambda_1} \right)^{p_{\max}} \left(1 + \frac{p_{\max}(p_{\max}+1)\lambda_1}{2} \right)^{p_{\max}} \\
&\leq \left(\frac{1}{\lambda_1} \right)^{p_{\max}} \underbrace{\frac{\left(1 + p_{\max}(p_{\max}+1)2^{-p_{\max}}\right)^{p_{\max}}}{\left(p_{\max}!\right)^2}}_{F(p_{\max})},
\end{aligned}$$

where the last inequality is obtained by using the definition of $p_{\max}$ in Lemma S1 to bound $\lambda_1/2 \leq 2^{-p_{\max}}$.

To complete the proof, we bound the function $F(p_{\max})$ from above. The numerator of $F(p_{\max})$ converges to 1 as $p_{\max} \to \infty$, as seen by taking its logarithm and bounding it:

$$\begin{aligned}
&p_{\max} \log\left(1 + p_{\max}(p_{\max}+1)2^{-p_{\max}}\right) \\
&\quad \leq p_{\max}^2 (p_{\max}+1) 2^{-p_{\max}} \to 0 \quad \text{as } p_{\max} \to \infty.
\end{aligned}$$

Thus $F(p_{\max})$ decreases to zero as $p_{\max}$ increases. Numerical evaluation shows that $F(p_{\max})$ attains a maximum value of 2 at $p_{\max} = 1$. □

Lemma S4. *The maximum number of DNF rules with a given signature $\mathbf{K} \in \Delta$ is bounded as*

$$\max_{\mathbf{K} \in \Delta} |\mathcal{C}(\mathbf{K})| < (2d)^{1/\lambda_1}.$$

Proof. The number of conjunctions of degree p is $\binom{d}{p} 2^p$, where the factor of 2^p is due to there being two choices of conditions on each of the p selected variables. The number of DNF rules with signature $\mathbf{K}$ is therefore

$$|\mathcal{C}(\mathbf{K})| = \prod_{p=1}^{p_{\max}} \binom{\binom{d}{p} 2^p}{K_p} < \prod_{p=1}^{p_{\max}} \frac{\left(\binom{d}{p} 2^p\right)^{K_p}}{K_p!}.$$

Taking logarithms, we obtain

$$\max_{\mathbf{K}\in\Delta} \log|\mathcal{C}(\mathbf{K})| < \max_{\mathbf{K}} \sum_{p=1}^{p_{\max}} K_p \log\left(\binom{d}{p}2^p\right) - \log(K_p!) \quad \text{s.t.} \sum_{p=1}^{p_{\max}} K_p(\lambda_0 + p\lambda_1) \le 1. \tag{S11}$$

For simplicity, we drop the nonlinear term $-\log(K_p!) \le 0$. The right-hand side of (S11) then becomes a maximization of a linear function over a simplex. The maximum value is given by

$$\max_{p=1,\ldots,p_{\max}} \frac{\log\left(\binom{d}{p}2^p\right)}{\lambda_0 + p\lambda_1} \tag{S12}$$

(attained by setting $K_{p^*} = 1/(\lambda_0 + p^*\lambda_1)$ for a maximizing value p^* and $K_p = 0$ otherwise). Again for simplicity, we further bound (S12) from above by dropping λ_0 from the denominator, resulting in

$$\max_{\mathbf{K}\in\Delta} \log|\mathcal{C}(\mathbf{K})| < \frac{1}{\lambda_1} \max_{p=1,\ldots,p_{\max}} \frac{1}{p}\log\binom{d}{p} + \log 2$$

(otherwise (S12) may require solving a transcendental equation). Since $\log\binom{d}{p}$ increases sublinearly with p, the maximum occurs at $p = 1$, yielding the desired result. □

By combining (S9), Lemmas S3 and S4, we obtain the desired bound on the number of DNF rules satisfying the optimality conditions in Lemmas S1 and S2.

Lemma S5. *Under the assumptions of Lemma S1, the number of DNF rules satisfying the necessary conditions of optimality in Lemmas S1 and S2 is bounded as*

$$|\mathcal{C}| < 2(2d)^{1/\lambda_1}\left(\frac{1}{\lambda_1}\right)^{p_{\max}}.$$

D.3 Proof of Theorem 1

We prove the theorem in two steps, first relating the empirical estimator in (7) to a problem intermediate between (5) and (7),

$$\mathcal{S}^* := \arg\min_{\mathcal{C}} \; Q(\mathcal{C}) := \bar{V}(\mathcal{C}) + R(\mathcal{C}) \quad \text{subject to} \sum_{i\in\mathcal{I}} \mathbb{1}[x_i \in \mathcal{C}] \ge \alpha m, \tag{S13}$$

and then relating this intermediate problem (S13) to (5). Problem (S13) has the same regularized volume objective as (5) but with the empirical probability constraint of (7).

For the first step, let $\hat{V}(\mathcal{C})$ denote the empirical volume in (7) (i.e. the first term in the objective function). As noted in Section 4.1, $\hat{V}(\mathcal{C})$ is a scaled binomial random variable with n trials and mean $\bar{V}(\mathcal{C})$. Hoeffding's inequality thus provides the following tail bound:

$$\Pr\left(\left|\hat{V}(\mathcal{C}) - \bar{V}(\mathcal{C})\right| > \epsilon_n\right) \le 2e^{-2n\epsilon_n^2}.$$

Defining $\hat{Q}(\mathcal{C}) = \hat{V}(\mathcal{C}) + R(\mathcal{C})$ and recalling that $Q(\mathcal{C}) = \bar{V}(\mathcal{C}) + R(\mathcal{C})$, the same bound holds for the difference $\hat{Q}(\mathcal{C}) - Q(\mathcal{C})$. Taking the union bound over the hypothesis class $\mathcal{C}$ yields

$$\Pr\left(\exists\mathcal{C}\in\mathcal{C} : \left|\hat{Q}(\mathcal{C}) - Q(\mathcal{C})\right| > \epsilon_n\right) \le 2|\mathcal{C}|e^{-2n\epsilon_n^2}. \tag{S14}$$

Assuming that the event in (S14) is not true, we obtain the following sequence of bounds, where the second inequality is due to the optimality of $\hat{\mathcal{S}}$ in (7):

$$Q(\hat{\mathcal{S}}) \le \hat{Q}(\hat{\mathcal{S}}) + \epsilon_n \le \hat{Q}(\mathcal{S}^*) + \epsilon_n \le Q(\mathcal{S}^*) + 2\epsilon_n. \tag{S15}$$

For this to hold with probability at least $1-\delta$, we set δ equal to the right-hand side of (S14) to obtain

$$\epsilon_n = \sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{2n}}. \tag{S16}$$

For the second step, we observe that the empirical probability $\hat{P}(\mathcal{C}) = \frac{1}{m}\sum_{i\in\mathcal{I}} \mathbb{1}[x_i \in \mathcal{C}]$ is also a scaled binomial random variable, this time with m trials and mean $P(\mathcal{C})$. We thus have a similar bound as in (S14),

$$\Pr\left(\exists\mathcal{C}\in\mathcal{C} : \left|\hat{P}(\mathcal{C}) - P(\mathcal{C})\right| > \epsilon_m\right) \le 2|\mathcal{C}|e^{-2m\epsilon_m^2},$$

and setting the right-hand side equal to δ yields the same expression for ϵ_m as in (S16) with n replaced by m. We then use Theorem 3 and Corollary 12 in (Scott and Nowak, 2006) to conclude that with probability at least $1-\delta$,

$$Q(\mathcal{S}^*) \le q^*(\alpha + \epsilon_m) \quad \text{and} \quad P(\mathcal{S}^*) \ge \alpha - \epsilon_m.$$

Indeed, since $\hat{S} \in \mathcal{C}$ and satisfies the constraint $\hat{P}(\hat{\mathcal{S}}) \ge \alpha$ as well, the above may be changed to

$$Q(\mathcal{S}^*) \le q^*(\alpha + \epsilon_m) \quad \text{and} \quad P(\hat{\mathcal{S}}) \ge \alpha - \epsilon_m. \tag{S17}$$

Combining (S15) and (S17) gives

$$Q(\hat{\mathcal{S}}) \le q^*(\alpha + \epsilon_m) + 2\epsilon_n \quad \text{and} \quad P(\hat{\mathcal{S}}) \ge \alpha - \epsilon_m$$

with probability at least $1-2\delta$.

Lastly, we use Lemma S5 to bound ϵ_n from above by

$$\sqrt{\frac{\lambda_1^{-1}\log(2d) + p_{\max}\log\lambda_1^{-1} + \log(4/\delta)}{2n}}$$

and similarly for ϵ_m.

E Generalization of the product estimator

Below, we give a Theorem bounding the expected error of the two-stage estimate $\hat{\mathcal{O}} = \hat{\mathcal{S}} \cap \hat{\mathcal{B}}$ as a function of the error of the base estimators $\hat{\mathcal{S}}, \tilde{\mathcal{B}}$. This justifies the two-stage nature of our algorithm and motivates selecting hyperparameters for overlap rules $\hat{\mathcal{B}}$ based on the error with respect to the base estimator $\tilde{\mathcal{B}}$. Before we state the result, we give a Lemma bounding the error of an estimator of a product of functions in terms of estimators of the respective terms in the product.

Consider the task of predicting the binary deterministic label $g(X) = g_1(X)g_2(X)$ by approximating the product of estimators f_1, f_2 of g_1, g_2. Now, let $R_g(f)$ denote the expected zero-one loss of f with respect to g over p,

$$R_g(f) = \mathop{\mathbb{E}}_{X \sim p} [\mathbb{1}[f(x) \neq g(x)]] \ .$$

Lemma S6. *For f_1 and f_2 such that $R_{g_1}(f_1) \leq A \leq \min\{p(f_2(X) = 1), p(g_2(X) = 1)\}$, $R_{g_2}(f_2) \leq B \leq \min\{p(f_1(X) = 1), p(g_1(X) = 1)\}$ and $\max\{A + B, C\} \leq 1/2$, let $f(X)$ approximate $f_1(X)f_2(X)$ and assume that $R_{f_1 f_2}(f) \leq C$. Then,*

$$R_g(f) \leq A + B + C$$

Proof. For convenience, let $f_1 = f_1(X), g_1 = g_1(X)$, et cetera, and let $\gamma = p(g(X) = 1)$.

$$\begin{aligned}
&R_g(f_1 f_2) = p(f_1 f_2 \neq g_1 g_2) \\
&= p(f_1 = f_2 = 1 \wedge (g_1 = 0 \vee g_2 = 0)) \\
&+ p((f_1 = 0 \vee f_2 = 0) \wedge g_1 = g_2 = 1) \\
&\leq p(f_1 = f_2 = 1 \wedge g_1 = 0) + p(f_1 = f_2 = 1 \wedge g_2 = 0) \\
&+ p(g_1 = g_2 = 1 \wedge f_1 = 0) + p(g_1 = g_2 = 1 \wedge f_2 = 0) \\
&\leq \min\{p(h_2 = 1), p(f_1 = 1 \wedge g_1 = 0)\} \\
&+ \min\{p(f_1 = 1), p(f_2 = 1 \wedge g_2 = 0)\} \\
&+ \min\{p(g_2 = 1), p(g_1 = 1 \wedge f_1 = 0)\} \\
&+ \min\{p(g_1 = 1), p(g_2 = 1 \wedge f_2 = 0)\} \\
&\leq A + B
\end{aligned}$$

In the first inequality, we use the standard Frechet inequalities. In the second and third, we use the assumptions in the statement. Alternatively, we could arrive at the same result by assuming that h_2 and (f_1, h_1) as well as h_1 and (f_2, h_2) are independent and decomposing the joint distributions. This could be guaranteed by sample splitting. We could then remove the assumption that the marginal probability of the label is larger than the error. In either case,

$$\begin{aligned}
R_g(f) &= p(f = f_1 f_2 \wedge f_1 f_2 \neq g) \\
&\quad + p(f \neq f_1 f_2 \wedge f_1 f_2 = g) \\
&\leq \min\{p(f = f_1 f_2), p(f_1 f_2 \neq g)\} \\
&\quad + \min\{p(f \neq f_1 f_2), p(f_1 f_2 = g)\} \\
&= p(f_1 f_2 \neq g) + p(f \neq f_1 f_2) \\
&\leq A + B + C \ .
\end{aligned}$$

□

We now state our result. First, we view membership in $\hat{\mathcal{O}} = \hat{\mathcal{S}} \cap \hat{\mathcal{B}}$ as given by an instance of the hypothesis class $\mathcal{F} = \{f(x) \coloneqq \mathbb{1}[x \in \hat{\mathcal{S}}] h(x); h \in \mathcal{H}\}$, for some function family $\mathcal{H}$. Then, let $R_g(f) = \mathbb{E}_{X \sim p}[\mathbb{1}[f(x) \neq g(x)]]$ denote the expected risk of f with respect to g over p, and $\hat{R}_g(f) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[f(x_i) \neq g(x_i)]$ the empirical risk.

Theorem S1. *Given are classifiers $\hat{s}, \tilde{b}$ of support membership s and propensity boundedness b, with overlap defined as $o(x) = s(x)b(x)$, such that for all $n > N$ it holds for $A_n, C_n \in \tilde{\mathcal{O}}(1/\sqrt{n})$ with $\max\{A_n, C_n\} \leq 1/4$ that $R_s(\hat{s}) \leq A_n, R_b(\tilde{b}) \leq C_n$. Then, for any function $\hat{o} \in \mathcal{H}$ approximating $\hat{s} \cdot \tilde{b}$, with probability larger than $1 - \delta$,*

$$R_o(\hat{o}) \leq \hat{R}_{\hat{s} \cdot \tilde{b}}(\hat{o}) + \frac{D_{\mathcal{F}, \delta, n}}{\sqrt{n}} + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right) \ ,$$

where $D_{\mathcal{F}, \delta, n} = \sqrt{8d(\log \frac{2m}{d} + 1) + 8 \log \frac{4}{\delta}}$, with d the VC-dimension of $\mathcal{F}$ and $\tilde{O}$ hides logarithmic factors.

Proof. From Lemma S6 and assumptions, we have that

$$R_o(\hat{o}) \leq R_{\hat{s} \cdot \tilde{b}}(\hat{o}) + R_s(\hat{s}) + R_b(\tilde{b}) \leq R_{\hat{s} \cdot \tilde{b}}(\hat{o}) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right) \ .$$

By applying standard VC-theory w.r.t. $\mathcal{F}$, we have our result. □

Theorem S1 bounds the generalization error of (e.g., Boolean rule) approximations of $\sqrt{n}$-consistent base estimators. It may be generalized to other rates, but convergence at *some* rate is necessary for consistency of the final estimator. Critically, the bias incurred by the approximation is observable and may be traded off for interpretability.

References

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.

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.

Kahn, H. (1955). Use of Different Monte Carlo Sampling Techniques. Technical report, RAND Corporation, Santa Monica, California.

Kallus, N. and Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 84:1243–1251.

Pearl, J. (2009). *Causality*. Cambridge university press.

Sanchez, G. V., Babiker, A., Master, R. N., Luu, T., Mathur, A., and Bordon, J. (2016). Antibiotic Resistance among Urinary Isolates from Female Outpatients in the United States in 2003 and 2012. *Antimicrobial Agents and Chemotherapy*, 60(5):2680–2683.

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

Shapiro, D. J., Hicks, L. A., Pavia, A. T., and Hersh, A. L. (2013). Antibiotic prescribing for adults in ambulatory care in the USA, 2007–09. *Journal of Antimicrobial Chemotherapy*, 69(1):234–240.

Sutton, R. S. and Barto, A. G. (2017). *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.

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)*.

Yau, S. T. and Zhang, L. (2006). An upper estimate of integral points in real simplices with an application to singularity theory. *Math. Res. Lett.*, 13(6):911–921.

Table S6: Population averages for covariates in Opioids in order of difference between the overlapping and non-overlapping set. DMME, MME and Duration are the medians of daily MME, total MME and prescription duration days in each group.

	Total	**DMME**	**MME**	**Duration**
Total sample	35106	46	225	5
Male	9301	50	300	5
Female	25805	45	225	5
Age groups				
<15	847	20	100	5
15-24	3334	45	200	5
25-34	9994	45	210	4
35-44	6820	46	225	5
45-54	6196	50	250	5
55-64	7915	50	300	5
>=65	0	0	0	0
Surgery type				
Auditory	29	18	135	6
Cardiovascular	3633	45	270	5
Integumentary	1507	48	225	5
Mediastinum	54	47	300	5
Female genital	3913	48	225	5
Hemic	885	50	225	5
Respiratory	665	45	250	5
Endocrine	214	45	200	5
Nervous	4350	60	375	6
Urinary	1476	45	225	5
Musculoskeletal	6678	60	450	7
Maternity	13553	45	200	4
Male genital	585	45	225	5
Year				
2011	7547	45	225	5
2012	10743	46	225	5
2013	9651	50	225	5
2014	7165	45	225	5
Diagnosis history (until day before surgery)				
Other specified gastritis: without mention of hemorrhage	491	42	225	5
Other ascites	233	45	225	5
Lumbosacral spondylosis without myelopathy	1135	60	400	6
Nausea with vomiting	1914	45	225	5
Other respiratory abnormalities	1935	45	225	5
Vomiting alone	765	45	200	5
Myalgia and myositis: unspecified	1522	50	250	5
Attention deficit disorder with hyperactivity	370	45	225	5
Attention deficit disorder without mention of hyperactivity	444	45	225	5
Depressive disorder: not elsewhere classified	2221	50	225	5
Dysthymic disorder	752	50	225	5
Tachycardia: unspecified	631	45	225	5
Degeneration of cervical intervertebral disc	904	56	337	6
Flatulence: eructation: and gas pain	427	45	225	5
Generalized anxiety disorder	833	45	225	5
Other symptoms referable to back	368	50	300	5
Cellulitis and abscess of leg: except foot	450	45	225	5
Constipation: unspecified	1136	45	225	5
Thoracic or lumbosacral neuritis or radiculitis: unspecified	1676	60	326	6
Anxiety state: unspecified	2205	50	225	5
Lumbago	4559	50	250	5
Abdominal pain: generalized	1607	45	225	5
Degeneration of lumbar or lumbosacral intervertebral disc	1542	60	388	6
Other and unspecified noninfectious gastroenteritis and colitis	1254	45	225	5
Major depressive affective disorder: recurrent episode: moderate	507	45	225	5
Asthma: unspecified type: unspecified	2044	45	225	5
Arthrodesis status	178	60	450	7
Chest pain: unspecified	4701	45	225	5

Routine general medical examination at a health care facility	9529	50	225	5
Diarrhea	1714	50	225	5
Fitting and adjustment of vascular catheter	318	45	225	5
Hypopotassemia	721	45	225	5
Bariatric surgery status	302	40	200	5
Sprain of neck	816	50	225	5
Unspecified gastritis and gastroduodenitis: without mention of hemorrhage	960	45	225	5
Injury of face and neck	271	46	300	5
Backache: unspecified	2471	50	225	5
Unspecified septicemia	222	45	225	5
Acute pharyngitis	4219	45	225	5
Acute bronchitis	3311	46	225	5
Abdominal pain: other specified site	2890	45	225	5
Atrophic gastritis: without mention of hemorrhage	537	45	225	5
Cough	3946	45	225	5
Altered mental status	202	45	225	5
Cervicalgia	2758	50	250	5
Abdominal pain: unspecified site	6339	45	225	5
Other chronic pain	346	56	300	6
Headache	3514	45	225	5
Tobacco use disorder	1834	50	225	5
Other screening mammogram	5722	50	240	5
Observation and evaluation for other specified suspected conditions	337	45	225	5
Unspecified sinusitis (chronic)	1624	46	225	5
Rheumatoid arthritis	353	50	300	5
Brachial neuritis or radiculitis NOS	1147	50	300	5
Loss of weight	455	46	225	5
Hypersomnia with sleep apnea: unspecified	424	42	225	5
Insomnia: unspecified	968	50	225	5
Other malaise and fatigue	5178	46	225	5
Other injury of chest wall	210	50	300	5
Dehydration	841	45	225	5
Acute respiratory failure	120	40	225	5

Table S7: Population averages for the 156 features in the UTI cohort. Mean values and total (for binary features) are given, and there are 64593 subjects in total.

	Mean	**Total**
Demographics		
Age	55.1	
Male	16.53%	10685
White	72.17%	46662
Veteran	4.61%	2981
Current Location		
Outpatient	64.89%	41957
Emergency Room	15.69%	10142
Inpatient	17.26%	11159
Intensive Care Unit (ICU)	2.69%	1736
Local Resistance Rates (Past 30-90 days, at this location)		
Trimethoprim/Sulfamethoxazole	18.61%	
Nitrofurantoin	19.85%	
Ciprofloxacin	22.70%	
Levofloxacin	24.19%	
Secondary Site of Infection		
Skin / Soft Tissue	0.20%	132
Blood	1.59%	1031
Respiratory Tract	0.53%	341
Nasal or Rectal Swab	0.19%	124
Medical History (Past 90 Days)		
Alcohol abuse	1.66%	1074
Deficiency anemia	2.84%	1837
Cardiac arrhythmias	17.08%	11041

Blood loss anemia	0.49%	315
Congestive heart failure	10.16%	6571
Coagulopathy	3.81%	2466
Diabetes, uncomplicated	14.13%	9135
Diabetes, complicated	5.00%	3232
Depression	11.80%	7627
Drug abuse	1.72%	1114
Fluid and electrolyte disorders	13.84%	8946
AIDS/HIV	0.43%	281
Hypertension, uncomplicated	32.51%	21017
Hypertension, complicated	5.43%	3513
Hypothyroidism	7.86%	5085
Liver disease	4.36%	2822
Lymphoma	1.63%	1051
Metastatic cancer	5.50%	3559
Other neurological disorders	6.68%	4319
Obesity	6.70%	4332
Pulmonary circulation disorders	3.13%	2025
Peptic ulcer disease, excluding bleeding	0.61%	393
Peripheral vascular disorders	5.68%	3672
Paralysis	3.08%	1992
Psychoses	2.42%	1563
Chronic pulmonary disease	11.29%	7299
Renal	8.87%	5735
Rheumatoid arthritis / collagen vascular diseases	3.76%	2428
Solid tumor without metastasis	12.00%	7760
Valvular disease	7.79%	5034
Weight loss	3.59%	2319
Preganant	3.08%	1989
Previous Care (Past 90 days)		
Inpatient Stay	18.38%	11882
Nursing Home Stay	1.20%	779
Previous Procedures (Past 90 days)		
Central Venous Catheder	5.27%	3410
Hemodialysis	0.66%	427
Mechanical Ventilation	5.74%	3714
Parenteral Nutrition	0.67%	434
Surgery	59.84%	38689
Previous Organisms (Past 90 days)		
Citrobacter species	0.42%	270
Coagulate negative Staphylococcus species	1.15%	741
Enterobacter aerogenes	0.15%	95
Escherichia coli	7.82%	5057
Enterococcus species	2.66%	1718
Enterobacter cloacae	0.29%	186
Group B Streptococcus	0.17%	109
Klebsiella pneumoniae	2.02%	1307
Morganella species	0.11%	73
Pseudomonas aeruginosa	0.92%	594
Proteus species	0.69%	445
Staph aureus	1.55%	1003
Serratia species	0.22%	145
Previous Resistance, measured by culture (Last 90 Days)		
Amoxicillin Clavulanate	2.34%	1511
Amikacin	0.10%	67
Ampicillin	7.44%	4808
Aztreonam	0.95%	616
Ceftazidime	0.30%	197
Cefazolin	9.22%	5962
Chlorampenicol	0.17%	111
Ciprofloxacin	4.62%	2984
Clindamycin	0.97%	624
Ceftriaxone	1.24%	804
Doxycycline	0.39%	249

Ertapenem	0.14%	88
Erythromycin	3.71%	2399
Cefepime	0.54%	351
Cefoxitin	0.49%	319
Gentamicin	1.65%	1066
Gentamicin (Synergistic)	0.47%	307
Imipenem	0.47%	303
Levofloxacin	5.32%	3439
Linezolid	0.09%	58
Meropenem	0.13%	85
Moxifloxacin	0.86%	556
Nalidixic Acid	0.09%	60
Nitrofurantoin	4.06%	2628
Oxacillin	1.79%	1158
Penicillin	2.41%	1559
Piperacillin	0.62%	402
Polymyxin B	1.22%	790
Rifampin	0.80%	518
Ampicillin Sulbactam	1.63%	1056
Streptomycin (Synergistic)	0.23%	150
Trimethoprim Sulfamethoxazole	3.10%	2006
Tetracycline	5.33%	3443
Ticarcillin	0.24%	153
Tobramycin	0.31%	203
Piperacillin Tazobactam	0.53%	341
Vancomycin	0.92%	598
Previous Antibiotic Prescription (Last 90 Days)		
Amikacin	0.09%	60
Amoxicillin	2.47%	1596
Amoxicillin/Clavulanate	2.15%	1388
Amphotericin B	0.16%	102
Ampicillin/Sulbactam	0.34%	217
Azithromycin	2.86%	1847
Aztreonam	0.25%	159
Cefadroxil	0.15%	96
Cefazolin	4.87%	3150
Cefepime	2.30%	1489
Cefixime	0.26%	166
Cefotetan	0.18%	114
Cefoxitin	0.25%	161
Cefpodoxime	0.88%	570
Ceftazidime	0.73%	475
Ceftriaxone	2.75%	1775
Cefuroxime	0.24%	156
Cephalexin	2.31%	1496
Ciprofloxacin	11.09%	7170
Clarithromycin	0.35%	226
Clindamycin	1.84%	1187
Daptomycin	0.10%	63
Dicloxacillin	0.19%	126
Doxycycline	1.73%	1119
Ertapenem	0.22%	140
Erythromycin	0.39%	249
Fluconazole	3.56%	2301
Fosfomycin	0.36%	232
Gentamicin	0.94%	607
Imipenem	0.33%	216
Levofloxacin	5.94%	3838
Linezolid	0.73%	470
Meropenem	0.40%	256
Metronidazole	4.49%	2906
Micafungin	0.24%	154
Minocycline	0.20%	129
Moxifloxacin	0.27%	174
Nafcillin	0.24%	157
Nitrofurantoin	2.73%	1767

Norfloxacin	4.25%	2749
Penicillin	0.31%	199
Piperacillin	0.41%	268
Piperacillin/Tazobactam	0.23%	148
Polymyxin B	0.52%	333
Posaconazole	0.18%	118
Tetracycline Metronidazole	0.09%	59
Trimethoprim	0.12%	79
Trimethoprim/Sulfamethoxazole	3.96%	2558
Vancomycin	8.80%	5690
Vancomycin Gentamicin	3.35%	2165