Maximizing Intervention Effectiveness

Vishal Gupta
Marshall School of Business, University of Southern California, guptavis@marshall.usc.edu

Brian Rongqing Han
Marshall School of Business, University of Southern California, rongqinh@marshall.usc.edu

Song-Hee Kim
Marshall School of Business, University of Southern California, songheek@marshall.usc.edu

Hyung Paek
Yale New Haven Hospital, hyung.paek@ynhh.org

Frequently, policymakers seek to roll out an intervention previously proven effective in a research study, perhaps subject to resource constraints. However, since different subpopulations may respond differently to the same treatment, there is no a priori guarantee that the intervention will be as effective in the targeted population as it was in the study. How then should policymakers target individuals to maximize intervention effectiveness? We propose a novel robust optimization approach that leverages evidence typically available in a published study. Our approach is tractable – real-world instances are easily optimized in minutes with off-the-shelf software – and flexible enough to accommodate a variety of resource and fairness constraints. We compare our approach with current practice by proving performance guarantees for both approaches, which emphasize their structural differences. We also prove an intuitive interpretation of our model in terms of regularization, penalizing differences in the demographic distribution between targeted individuals and the study population. Although the precise penalty depends on the choice of uncertainty set, we show that for special cases we can recover classical penalties from the covariate matching literature on causal inference. Finally, using real data from a large teaching hospital, we compare our approach to common practice in the particular context of reducing emergency department utilization by Medicaid patients through case management. We find that our approach can offer significant benefits over common practice, particularly when the heterogeneity in patient response to the treatment is large.

Key words: analytics, robust optimization, intervention effectiveness, healthcare

1. Introduction

Across domains, we observe an increasingly common decision-making paradigm: Researchers assess whether a particular intervention is effective in treating a study population, such as in a randomized control trial. Practitioners then roll out successful interventions to a potentially different candidate population, hoping to achieve similar effectiveness. Although this paradigm is perhaps most familiar in medicine, "intervention" and "treatment" can be interpreted quite generally, as these terms

might refer to an after-school training program to reduce childhood obesity (Vizcaíno et al. 2008) or a tuition reduction program to increase college enrollment (Deming and Dynarski 2009).

Despite its ubiquity, this paradigm faces practical challenges. Many interventions are too expensive to provide to everyone in the candidate population (see, e.g., our case study below). Thus, practitioners must solve a resource allocation problem: Who should be targeted for treatment? A first intuition, motivated by medical practice, might be to target the "sickest patients," i.e., the individuals most in need of treatment. However, these sickest patients may be too sick to benefit from treatment, so targeting them is arguably an inefficient use of resources. More poignantly, different individuals may respond differently to the same intervention. A prudent decision-maker would ideally target those individuals who benefit the most in order to maximize the aggregate benefit subject to resource constraints. The challenge, of course, is identifying the potential benefit for *each* individual prior to administering the intervention.

In this paper, we propose a robust optimization approach to this resource allocation problem using only the evidence typically *published* in a research study, including its inclusion/exclusion criteria, the demographic features of its study population and estimates of the average benefit. We provide a precise specification of this evidence in Section 2.2, but note that it does *not* usually include the raw, individual-level study data.

This restriction to the published evidence is a key distinguishing feature of our work. There is a growing body of literature on predicting the potential benefit of treatment at an individual level, i.e., learning a heterogeneous causal effect, for either estimation or personalization (see, e.g., Imai et al. 2013, Athey and Imbens 2015, Kallus 2017 and references therein). In the marketing literature, these methods are sometimes referred to as "uplift modeling" (Gutierrez and Gérardy 2017, Zhao et al. 2017, Ascarza 2018). There is also a second stream of literature on estimating the average causal effect in a population distinct from the original study population (e.g., Cole and Stuart 2010, Stuart et al. 2011, Hartman et al. 2015). In principle, either approach might be adapted to solve the above resource allocation problem.

However, these methods typically require access to individual-level data from patients in the study. Such data are rarely available in practice (Eichler et al. 2012). Indeed, unlike the typical marketing and personalization settings, the decision-makers rolling out interventions in healthcare and policy-making contexts are often distinct from the researchers studying those interventions. Moreover, laws such as the Health Insurance Privacy and Portability Act (HIPPA), the General Data Protection Regulation (GDPR), and Family Educational Rights and Privacy Act (FERPA) heavily restrict those researchers from sharing the raw study data with policy makers out of patient-privacy and ethical concerns. Without this raw study data, it is not possible to implement the above approaches. Worse, even when study data are available, they are frequently inadequate for

the task. Since conducting randomized control trials is notoriously expensive, most studies are sized to have just enough power to detect an average causal effect but are typically too small to learn the precise heterogeneous effects across patients. Learning this heterogenous effect is critical to the above resource allocation problem.

Since learning heterogeneous effects without the raw study data seems impractical, we take a different approach via robust optimization (Ben-Tal et al. 2009). Generally, robust optimization methods optimize worst-case performance over an uncertainty set of possible realizations of uncertain parameters. Our particular robust approach seeks the subset of patients that maximizes the worst-case aggregate intervention effectiveness, where the worst-case is taken over an uncertainty set of models for the heterogeneous causal effect that are consistent with the published study evidence. In other words, rather than learning a single model for causal effects, we optimize the targeting to ensure the effect over many plausible models is as large as possible. Depending on the precise assumptions on the set of models, this optimization problem can be cast as a mixed-binary linear optimization problem or a mixed-binary second-order cone optimization problem, both of which can be readily solved with off-the-shelf software. The resulting formulations are also flexible enough to easily accommodate side constraints on the targeting, such as budget, operational or fairness constraints, and to incorporate evidence from multiple papers.

We prove that our robust approach is equivalent to approximating the targeted subset's average effectiveness by the study population's average effectiveness *minus* a penalty that depends on the differences in demographics between these two groups, an insight that we term "covariate matching as regularization." Intuitively, the more different the targeted subset and study population are, the less accurately the study population's effectiveness approximates the targeted subset's effectiveness. The precise form of the penalty depends on the particular uncertainty set. For special cases, we show that the penalty coincides with common techniques used for covariate matching in causal inference – χ^2-matching, mean matching, and Mahalanobis matching – highlighting an interesting theoretical connection between these two areas. (Kallus 2016 observes a similar connection in the context of designing experiments.)

We stress that our robust approach does not directly estimate individual-level causal effects, but only approximates the aggregate effect over the targeted subset. This "portfolio" viewpoint sharply contrasts with both the aforementioned statistical literature and current practice. Indeed, most common approaches to targeting employ so-called *scoring rules*: Practitioners assign each individual in the candidate population a score approximating her unknown heterogeneous causal effect and then target individuals with the highest scores. Scores are typically informed by some combination of domain expertise and predictive modeling. However, recent empirical studies have called the effectiveness of such rules into question (Jackson and DuBard 2015). In Section 3, we

provide a theoretical analysis of scoring rules, providing sufficient conditions for their optimality and a tight performance guarantee when they are suboptimal. In particular, we prove that if patients may experience adverse effects from the particular treatment, scoring rules may perform arbitrarily badly and may be worse than not providing treatment to anyone.

This research was inspired by our partner hospital, which seeks to reduce excessive emergency department (ED) utilization by adult Medicaid patients by rolling out a case-management intervention. (See Section 1.2 for details.) We use real data from this case study as a running example and to assess our methodology (Section 5). We summarize our work as follows:

1. We formalize an optimization approach to maximize intervention effectiveness using the evidence typically available in a published study. To the best of our knowledge, we are the first to address this problem using only published study data.
2. We prove tight performance bounds for current practice (scoring rules). In particular, we prove that scoring rules can perform arbitrarily badly when the treatment is potentially harmful.
3. We propose a robust optimization approach to maximize the worst-case performance over a large class of models that agree with the study evidence. To the best of our knowledge, we are the first to apply robust optimization methods to the analysis of summary-level causal inference data. Our model is flexible enough to accommodate a variety of side constraints and can be solved for real-world instances within a few minutes using commercial software. Moreover, its worst-case performance is bounded by a value that depends on the class of models and the true heterogeneous effect. Under some mild assumptions, this constant is zero, ensuring that our robust approach is never worse than not targeting, even when the treatment could be potentially harmful.
4. We provide an intuitive interpretation of our robust model as "covariate matching as regularization," connecting with the literature on causal inference and illustrating how canonical covariate matching techniques can be recovered as special cases.
5. Using data from our partner hospital, we show that our robust approach performs almost as well as scoring rules when the degree of heterogeneity in causal effects is small and can perform much better than scoring rules as the degree of heterogeneity increases, especially when the treatment is potentially harmful.

1.1. Connections to Existing Literature

Our work connects to a growing body of robust optimization applications, particularly in healthcare operations (e.g., Bortfeld et al. 2008, Deo et al. 2015, Chan et al. 2016, 2017, Goh et al. 2018.) Adopting a worst-case perspective is appealing, particularly in healthcare, where decision-makers aspire to "do no harm", and high-costs and consequences fuel risk aversion. A distinguishing

feature of our work is that while many robust optimization models seek to immunize solutions against parameter uncertainty or implementation uncertainty, our formulation is more naturally interpreted as immunizing solutions against model uncertainty, i.e., our uncertainty about the true model for heterogeneous causal effects. In this respect, we are most similar to Bertsimas et al. (2016b). At the same time, we contribute to a large body of work connecting robust optimization and regularization (Ghaoui and Lebret 1997, Xu et al. 2009, Lam 2016, Bertsimas and Copenhaver 2017, Gao et al. 2017). In particular, our work elucidates the connection between the form of regularizer and assumed structure of the causal effects. This perspective, we feel, helps provide an alternate, statistical interpretation of common regularizers and uncertainty sets.

Moreover, our restriction to the published study evidence distinguishes our work from existing techniques in data-driven robust optimization. In particular, typical data-driven robust optimization models (e.g., Delage and Ye 2010, Esfahani and Kuhn 2015, Bertsimas et al. 2017, 2018) assume that the data are noisy versions of the underlying parameters or a sequence of i.i.d. realizations of random variables depending on those parameters. In our setting, such data would correspond to noisy observations of the potential causal effect in the candidate population *before* administering treatment. However, in causal inference settings, it is impossible to observe this effect directly (even noisily), a phenomenon sometimes called the *Fundamental Problem of Causal Inference* (see Holland 1986 or our discussion in Section 2). Worse, in our setting of interest, the data we do have pertains to a different population, the study population. These features are intrinsic to our application and require new modeling and methods.

In focusing on the published evidence, our work also connects to meta-regression techniques in statistics that aim to "pool" the results of different published studies to form a refined estimate of causal effects (Higgins and Thompson 2002, Bertsimas et al. 2016a). We differ from these works in two important respects: First, these methods' successes rely upon access to multiple distinct papers; it is by combining distinct sources of information that they refine estimates. By contrast, although our robust approach can be applied when multiple published studies are available, it applies equally well with only a single study. Second, and more critically, these methods generally focus on estimation and inference, not on decision-making. A notable exception is Bertsimas et al. (2016a), which does consider an optimization problem, but in a different context – designing a clinical trial versus rolling out an intervention – and with a different mathematical structure.

Finally, we contrast our work with the reinforcement learning literature. An alternate approach to rolling out an intervention might be to proceed sequentially, treating individuals in the candidate population one at a time, observing their response to treatment, and using those observations to decide whom to treat next. This approach is naturally modeled as a contextual multi-arm bandit problem (Bastani and Bayati 2016, Negoescu et al. 2017). While reinforcement learning is

a reasonable strategy for some interventions, for many others, the outcome of interest may take a long time to observe. For example, it may take months to check for a reduction in the childhood obesity rate. With this time delay, online approaches that proceed sequentially may be impractical, motivating our off-line treatment of the problem.

1.2. Case Study: Emergency Department Visits by Adult Medicaid Patients

Medicaid is a public insurance program for low-income, disabled and needy people under the age of 65. At our partner hospital, a large teaching hospital, approximately 50% of all ED visits are from Medicaid patients. Medicaid typically pays 50% less than private insurance(Zuckerman et al. 2009). Consequently, our partner hospital is underpaid for each Medicaid patient's ED visit. On the other hand, Medicaid patients generally suffer from multiple chronic diseases and lack easy access to primary care (Billings and Raven 2013). This combination of financial burden and patient need has sparked interest in intervention programs that might reduce unnecessary Medicaid ED visits while improving patients' health outcomes.

Case management is the most widely used intervention in reducing ED visits. While the implementation details differ between studies, at a high level, case management involves a team of social workers, nurses, and physicians providing crisis intervention, supportive therapy by phone or in person, referral to substance abuse services, linkage to primary care providers and assistance with making appointments to outpatient care. This team of professionals may also liaise with other assistance programs on the patient's behalf, such as to find subsidized housing. Prior research has shown case management to be effective in specific populations regarding reducing ED visits and improving patient outcomes (Shumway et al. 2008, Shah et al. 2011).

Unsurprisingly, case management is expensive, both financially and in terms of resources. Limited availability of physicians, nurses, social workers and psychiatrists prevents enrollment of all Medicaid patients in the program at our partner hospital. Thus, our case study seeks to use data to target a subset of adult Medicaid ED patients for case management to reduce ED utilization and underpayments while maintaining quality of care for this population. Based on their resource constraints, our partner hospital would ideally like to target approximately 200 patients.

2. Model Setup

2.1. Candidate Population

We seek to target at most $K > 0$ patients for intervention from a candidate population of size $C > K$ in order to maximize total intervention effectiveness. We adopt a potential outcome framework for causal inference (Imbens and Rubin 2015). For each patient $c \in \{1, \ldots, C\}$, there exists a *fixed* tuple $(\mathbf{x}_c, y_c(0), y_c(1), r_c)$. The parameters $\mathbf{x}_c$ and r_c are assumed *known*, while $y_c(0)$ and $y_c(1)$ are

unknown and represent potential outcomes. Specifically, $\mathbf{x}_c \in \mathcal{X}$ denotes patient c's pre-treatment covariates, e.g., demographic characteristics, and may include discrete and continuous components.

The quantity $y_c(0)$ (resp. $y_c(1)$) represents the outcome of interest for patient c if she does not (resp. does) receive the treatment. Before choosing whether to administer treatment, we know neither $y_c(0)$ nor $y_c(1)$. After this choice, *exactly one* of $y_c(0)$ and $y_c(1)$ is revealed depending on on our choice for patient c.

In our case study in Section 5, $y_c(0)$ and $y_c(1)$ denote the number of times patient c visits the ED in the next 6 months. In particular, smaller values are better. Consequently, we adopt a non-standard convention and define the *causal effect/treatment effect*[1] of patient c as $\delta_c \equiv y_c(0) - y_c(1)$. (It is usually the negative of this quantity.) We stress that δ_c may be positive or negative, i.e., the treatment may benefit or harm an individual patient. Moreover, since we never observe both $y_c(0)$ and $y_c(1)$, we cannot observe δ_c, directly, not even noisily.

We define the *intervention effectiveness* of patient c to be $r_c\delta_c$, where $r_c \geq 0$ represents a known reward. Adopting a linear model for effectiveness is with some loss of generality. However, we believe this model is a good approximation for our case study (see below) and many other applications. In many medical applications, one is not interested in a monetary outcome, but simply the aggregate benefit (in units of δ_c) across patients. In these cases, one can take $r_c = 1$ for all c. We stress that r_c may differ by patient and might depend in a complex way on the covariates x_c. For example, it might be the output of a machine-learning model that given x_c predicts (dollar) savings for each unit decrease in the outcome. In what follows, we assume without loss of generality that r_c is one of the components of $\mathbf{x}_c$ since both are known before targeting.

We seek to maximize the total intervention effectiveness as follows:

$$\max_{\mathbf{z}\in\mathcal{Z}} \sum_{c=1}^{C} z_c r_c \delta_c, \text{ where } \mathcal{Z} \equiv \left\{ \mathbf{z} \in \{0,1\}^C \;\middle|\; \sum_{c=1}^{C} z_c \leq K \right\}. \tag{1}$$

If we were to observe the causal effects δ_c directly, the optimal solution would be to rank each patient based on the intervention effectiveness $r_c\delta_c$ and to target the top K patients with non-negative values. Let $B^* \subseteq \{1,\ldots,C\}$ denote this solution, which we call the *full-information benchmark*. The challenge is that since we only observe one of $y_c(0)$ or $y_c(1)$ depending on our treatment assignment, we cannot observe the causal effects δ_c directly.

2.2. Study Population and Evidence for Treatment

Although we cannot learn δ_c, we will assume that we have some evidence from a published paper that the treatment is effective, namely, a confidence interval for the average causal effect in a study population and summary statistics for the pre-treatment covariates of that study population.

[1] We will use the term "causal effect" instead of "treatment effect" to distinguish from "intervention effectiveness" defined below.

Formally, let $(\mathbf{x}^s, y^s(0), y^s(1))$ for $s \in \{1, \ldots, S\}$ be the pre-treatment covariates and potential outcomes for each patient in the study population.[2] In general, the study and candidate populations may be distinct. The parameters $(\mathbf{x}^s, y^s(0), y^s(1))$ are fixed but unknown. Instead, we assume we know an interval $[\underline{I}, \overline{I}]$ such that $\underline{I} \leq \frac{1}{S}\sum_{s=1}^{S} \delta^s \leq \overline{I}$, where $\delta^s \equiv y^s(0) - y^s(1)$, for all $s \in \{1, \ldots, S\}$. The quantity $\frac{1}{S}\sum_{s=1}^{S} \delta^s$ is the Sample Average Treatment Effect (SATE).

Knowing $[\underline{I}, \overline{I}]$ is a mild assumption. Most studies, regardless of their precise statistical methodologies, report a confidence interval for SATE that can be used for $[\underline{I}, \overline{I}]$. For example, in randomized control trials (the gold standard for medical research), a simple t-test, a linear regression including pre-treatment covariates and the treatment assignment, or a matching estimator yields a confidence interval for SATE (see, e.g., Imbens 2004).

There do exist studies that do not report a confidence interval for SATE because of their chosen statistical design, such as a stratified analysis, which instead estimates average causal effects in each stratum. In our opinion, however, such designs are less common in healthcare. In special cases, we can still approximate a confidence interval for SATE for these studies (see, e.g., Section 2.3).

The interval $[\underline{I}, \overline{I}]$ tells us nothing about the distribution of $\mathbf{x}^s$ in the study population. Most studies therefore also report summary statistics for $\mathbf{x}^s$ and detailed inclusion/exclusion criteria. The precise statistics used (mean, median, standard deviation, etc.) often differ between studies.

To provide a flexible modeling framework for summary statistics, we assume that the study reports a set of *description functions* $\phi_g : \mathcal{X} \mapsto \mathbb{R}$, $g = 1, \ldots, G$ and their expectations over the study population, i.e., $\mu_g \equiv \frac{1}{S}\sum_{s=1}^{S} \phi_g(\mathbf{x}^s)$. By suitably choosing the functions ϕ_g (according to the study paper), we can model a wide variety of possible summary statistics as generalized moments of the study population's covariate distribution. In our opinion, most studies present a combination of summary statistics of the following three types:

Partition Description Functions: When there exists a natural partition of $\mathcal{X} = \bigcup_{i=1}^{I} \mathcal{X}_i$, e.g., patient race, studies often report the proportion of *type* i patients $\mu_i = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\mathbf{x}^s \in \mathcal{X}_i)$ for $i = 1, \ldots, I$. We model these statistics with description functions $\phi_i(\mathbf{x}) \equiv \mathbb{I}(\mathbf{x} \in \mathcal{X}_i)$ for $i = 1, \ldots, I-1$. Note that since $\mu_I = 1 - \sum_{i=1}^{I-1} \mu_i$ and $\phi_I(\mathbf{x}) = 1 - \sum_{i=1}^{I-1} \phi_i(\mathbf{x})$, it suffices to only specify these $I-1$ description functions to capture all I statistics. We assume for simplicity that $\mu_i > 0$ for $i = 1, \ldots, I$.

Linear Description Functions: When $\mathcal{X} \subseteq \mathbb{R}^I$ contains continuous variables, studies often report their mean values in the study population $\mu_i = \frac{1}{S}\sum_{s=1}^{S} x_i^s$ for all $i = 1, \ldots, I$. We model these statistics with description functions $\phi_i(\mathbf{x}) \equiv x_i$ for $i = 1, \ldots, I$.

[2] Note the distinction between superscripts and subscripts for $\mathbf{x}$. The value $\mathbf{x}^1$ describes the first patient in the study population, while $\mathbf{x}_1$ describes the first patient in the candidate population.

Quadratic Description Functions: When $\mathcal{X} \subseteq \mathbb{R}^I$, studies may report, in addition to the mean $m_i \equiv \frac{1}{S}\sum_{s=1}^{S} x_i^s$, the standard deviation $\sigma_i^2 \equiv \frac{1}{S}\sum_{s=1}^{S}(x_i^s - m_i)^2$ of each covariate for all $i = 1, \ldots, I$. We model the mean m_i with the I description functions above and the standard deviation with additional I description functions: $\phi_{I+i}(\mathbf{x}) \equiv x_i^2$ and $\mu_{I+i} = m_i^2 + \sigma_i^2$ for all $i = 1, \ldots, I$.

We stress that these data – summary statistics and SATE in a separate population – strongly contrast with the data typically available in data-driven robust optimization models. Indeed, traditional data-driven models often assume we observe (noisy) realizations of δ_c which, as mentioned, is impossible in our setting. Nonetheless, we will combine the description functions, their statistics and the study SATE in Section 4 to formulate our robust optimization model.

2.3. Case Study: Setup

The candidate population in our partner hospital is of size $C = 951$. (We defer a detailed description until Section 5.1.) Let $y_c(0)$ and $y_c(1)$ be the number of ED visits in the next 6 months if patient c in the candidate population does not or does, respectively, receive case management, Define $y^s(0)$ and $y^s(1)$ similarly for the study population. The unknown causal effect δ_c is the potential number of ED visits reduced by case management for patient c. Finally, let the known reward r_c be an estimate of the average charges per ED visit for patient c based on their medical history.

We assume a linear model for effectiveness for this application. ED charges typically consist of a relatively large fixed component common to most visits and a more variable idiosyncratic component that differs between visits. The fixed component corresponds to charges for doctor and staff time, basic equipment, and routine testing. The variable component corresponds to the additional services for the specific complaint on that visit and can be large for very sick patients. Consequently, charges are highly concentrated around the fixed component with a long tail. (See Fig. EC.1 in the e-companion.)

Intuitively, case management is unlikely to prevent visits corresponding to extreme medical events (e.g., strokes, falls among the elderly), i.e. the visits with high variable costs. Rather, case management might help prevent "less serious" visits (e.g., a person with dehydration from chronic malnutrition) whose costs are closer to the fixed cost (Billings et al. 2000). Thus, we approximate the marginal benefit of reducing 1 visit as a constant that may depend on the patient's covariates.

We use data from Shumway et al. (2008) as the study evidence because their study population mainly consists of low-income patients with behavioral problems who are similar to the Medicaid population in our hospital. Shumway et al. (2008) investigate the causal effects of case management in reducing ED visits among ED frequent users at San Francisco General Hospital, an urban public hospital. Patients were eligible for study participation if they had at least 5 visits to the ED in the

Table 1 Evidence of Causal Effects for Case Management from Shumway et al. (2008) for the Study Population

	Stratum 1 No. of ED Visits 5 - 11†	**Stratum 2 No. of ED Visits ≥ 12**	**Total**
No. of Patients			
Treatment	81 (32%)	86 (34%)	167 (66%)
Control	40 (15%)	45 (18%)	85 (34%)
No. of ED Visits in 6 Months			
Treatment, mean ± sd	2.5 ± 3.2	5.2 ± 5.6	3.9 ± 2.0*
Control, mean ± sd	4.6 ± 6.2	8.5 ± 9.6	6.7 ± 8.2*
CATE****, mean ± sd, [95% CI]**	2.1 ± 1.0, [0.1, 4.1]	3.3 ± 1.6, [0.2, 6.4]	
SATE, mean ± sd, [95% CI]			2.7 ± 1.0‡, [0.8, 4.6]

Notes. † Patients are stratified based on number of ED visits in the previous year. * Approximated by taking the weighted average of the mean and variance for each group. For example, the mean outcome for the treatment group is $(81 \times 2.5 + 86 \times 5.2)/(81 + 86) = 3.9$ ED visits. ** CATE refers to the Conditional Average Treatment Effect within each stratum. We formally define CATE in Section 4.1. ‡ The standard deviation is approximated as $\sqrt{s_1^2/n_1 + s_2^2/n_2}$, where s_i is the standard deviation of the outcome and n_i is the number of people in group i for $i = 1, 2$.

previous year, were San Francisco residents, were at least 18 years old and had psychosocial problems that might be addressed with case management. Such problems include housing problems, medical care problems, substance abuse, and mental health disorders. The study was conducted between 1997 and 1999, and a total of $S = 252$ eligible patients were enrolled. The authors performed a stratified analysis on two strata based on previous ED visits. We reproduce their results in Table 1 and summary statistics in Table 2 for the 252 patients. Since Shumway et al. (2008) do not directly present a confidence interval for SATE, we approximate it (see notes in Table 1 and also discussion at the beginning of Section 5).

Table 2 Summary Statistics for the Study and Candidate Populations

	Study Population ($S = 252$ patients)	**Candidate Population ($C = 951$ patients)**
Male	188 (75%)	442 (46%)
Race/Ethnicity		
African American	138 (54%)	475 (50%)
Hispanic	55 (22%)	7 (1%)
White	34 (13%)	283 (29%)
Other	28 (11%)	186 (20%)
Age, mean ± sd	43.3 ± 9.5	38.3 ± 12.5
No. of ED Visits in Previous Year*		
5 - 11	121 (48%)	860 (90%)
≥ 12	131 (52%)	91 (10%)
Most Frequent Diagnosis† during ED Visits	mental disorder (22%)	alcohol-related disorders (10%)
	injury (16%)	abdominal pain (6%)
	skin diseases (8%)	back problems (5%)
	endocrine discorders (5%)	nonspecific chest pain (4%)
	digestion disorders (5%)	connective tissue diseases (3%)
	respiratory illnesses (5%)	non-traumatic joint disorders (3%)

Notes. * Both populations only include patients who have had at least 5 ED visits. † Calculated from primary the ICD-10-CM diagnosis code using Clinical Classification Software (Elixhauser et al. 2014).

The second column of Table 2 presents the same summary statistics for the $C = 951$ Medicaid patients at our partner hospital who satisfy the inclusion/exclusion criteria of Shumway et al. (2008). Despite these criteria, these two populations still display some systematic differences.

Finally, we map Table 2 into our framework with the following description functions:

- The proportion of male patients can be represented as a partition description function with an indicator for gender. Race and whether the number of ED visits in the previous year exceeds 11 can also be represented with indicators.
- The average age can be represented as a linear description function.
- The standard deviation of age can be represented by a quadratic description function with corresponding summary statistics $43.3^2 + 9.5^2$.

3. Scoring Rules

Given the structure of the optimal solution to Problem (1), a natural heuristic is to approximate the unknown causal effect δ_c with some observable proxy $\hat{\delta}_c$ and then rank patients accordingly.

DEFINITION 1. Given a proxy $\hat{\delta}_c > 0$, the *$r\hat{\delta}$-scoring rule* ranks each patient c in the candidate population by $r_c\hat{\delta}_c$ and targets the K highest-ranked patients with non-negative scores.

In principle, one can use any observable metric as the proxy. We focus on two proxies that correspond to common assumptions and methods employed by practitioners:

Constant Effect Sizes and Reward Scoring: If one believes that the true causal effect is constant, i.e., $\delta_c = \delta_0 > 0$ for all $c \in \{1, \ldots, C\}$, then no matter what the value of δ_0 is, using the proxy $\hat{\delta}_c = 1$ and ranking patients by r_c (i.e., reward scoring, or r-scoring) yields an optimal solution to Problem (1). The assumption of constant effect sizes is common in statistical inference for randomized control trials. In particular, the most common approach for estimating the sampling variance of the SATE estimator assumes that the causal effects are constant across all individuals in the study population (Imbens 2004).

Proportional Effect Sizes and Outcome Scoring: If one believes that the true causal effect is proportional to the outcome without treatment, i.e., $\delta_c = \alpha y_c(0)$ for all $c \in \{1, \ldots, C\}$ and some $\alpha > 0$, then no matter what the value of α is, using the proxy $\hat{\delta}_c = y_c(0)$ and ranking patients by $r_c y_c(0)$ (i.e., outcome scoring, or $ry(0)$-scoring) yields an optimal solution to Problem (1).

In words, outcome scoring targets *high-risk* patients. Many studies have developed statistical or machine-learning models to predict the number of ED visits by a specific patient in the near future (Billings and Raven 2013). They estimate $y_c(0)$ (i.e., number of ED visits for patient c) and suggest focusing on patients with large estimates. Implicitly, such recommendations assume that the true causal effect δ_c is proportional to $y_c(0)$.

Neither reward scoring nor outcome scoring leverages the summary statistics for the study population but may leverage the candidate population covariates in a sophisticated way to estimate r_c

and $y_c(0)$. In the rest of this section, we show that scoring rules may be highly suboptimal when these underlying assumptions about the causal effects are violated.

3.1. Performance of Scoring Rules with Benign Treatment

The performance of scoring rules depends heavily on whether the treatment might be harmful. If we assume that we can identify and avoid treating patients with potential adverse events to that all treated patients experience positive causal effects, then scoring rules may perform very well.

THEOREM 1 **(Worst-Case Performance of Scoring Rules with Benign Treatment)**. *Without loss of generality, index patients so that* $r_1\hat{\delta}_1 \geq \cdots \geq r_C\hat{\delta}_C \geq 0$. *Suppose* $K \leq C/2$ *and there exists* $0 < \underline{\delta} < \overline{\delta} < \infty$ *such that* $\delta_c/\hat{\delta}_c \geq \underline{\delta} > 0$, *for all* $c \in \{1,\ldots,K\}$, *and* $\delta_c/\hat{\delta}_c \leq \overline{\delta}$, *where* $\overline{\delta} > 0$ *for all* $c \in \{K+1,\ldots,C\}$. *Then, the* $r\hat{\delta}$*-scoring rule obtains at least* $\omega(\underline{\delta}/\overline{\delta})$ *of the full-information benchmark optimal value, where*

$$\omega(\underline{\delta}/\overline{\delta}) \equiv \frac{(\underline{\delta}/\overline{\delta})\sum_{c=1}^{K} r_c\hat{\delta}_c}{(\underline{\delta}/\overline{\delta})\sum_{c=1}^{k^*} r_c\hat{\delta}_c + \sum_{c=K+1}^{2K-k^*} r_c\hat{\delta}_c} \tag{2}$$

and

$$k^* = \begin{cases} 0, & \text{if } (\underline{\delta}/\overline{\delta}) \leq r_{2K}\hat{\delta}_{2K}/r_1\hat{\delta}_1 \\ \arg\max\{c \mid 1 \leq c \leq K,\ (\underline{\delta}/\overline{\delta}) \geq r_{2K-c+1}\hat{\delta}_{2K-c+1}/r_c\hat{\delta}_c\}, & \text{otherwise.} \end{cases}$$

Moreover, for a given value of $\mathbf{r}$ *and* $\hat{\boldsymbol{\delta}}$, *there exist values of* $\boldsymbol{\delta}$ *such that the bound is tight.*

Intuitively, $\overline{\delta}$ measures how much we may have underestimated the causal effects of patients that were not picked, while $\underline{\delta}$ measures how much we may have overestimated the causal effect of patients that were picked. With this interpretation, the critical assumption in Theorem 1 is that $\delta_c/\hat{\delta}_c \geq \underline{\delta} > 0$, for all $c \in \{1,\ldots,K\}$. Indeed, since $r_c \geq 0$, this implies that $r_c\delta_c \geq 0$ for all $c \in \{1,\ldots,K\}$, i.e., targeted patients do not experience adverse effects.

At first reading, the function $\omega(\cdot)$ in Theorem 1 appears quite complicated. Importantly, it depends on the unknown causal effects only through the ratio $\underline{\delta}/\overline{\delta}$. Intuitively, this ratio measures the *degree of correspondence* between the proxy $\hat{\delta}_c$ and the true causal effect δ_c. If the proxy is reasonably accurate, i.e., $\hat{\delta}_c \approx \delta_c$ for all $c \in \{1,\ldots,C\}$, then we would expect $\underline{\delta} \approx \overline{\delta}$, and the ratio is close to 1. At the other extreme, if $\hat{\delta}_c$ and δ_c are very different for patients $\{1,\ldots,K\}$ compared to patients $\{K+1,\ldots,C\}$, the ratio $\underline{\delta}/\overline{\delta}$ will be close to 0. Intuitively, scoring rules should improve as the degree of correspondence increases. We prove that our bound $\omega(\cdot)$ shares these features:

COROLLARY 1. *Under the assumptions of Theorem 1,*

(1) $\omega(\underline{\delta}/\overline{\delta})$ *is increasing in* $\underline{\delta}/\overline{\delta}$*;*

(2) If $\underline{\delta}/\overline{\delta} \geq r_{K+1}\hat{\delta}_{K+1}/r_K\hat{\delta}_K$, $\omega(\underline{\delta}/\overline{\delta}) = 1$ *and the* $r\hat{\delta}$*-scoring rule is optimal;*

(3) $\omega(\underline{\delta}/\overline{\delta}) \to 0$ *as* $\underline{\delta}/\overline{\delta} \to 0$.

A numerical example showing typical values of the bound and its shape is in Section 3.3.

3.2. Performance of Scoring Rules with Potentially Harmful Treatment

Assuming that we can identify and avoid treating patients that would have an adverse response to the treatment is particularly strong; in practice, the treatment may be ineffective or even harmful. The evidence for case management, in particular, is mixed. Lee and Davenport (2006) provided case management to patients with over 3 ED visits in the previous month and found no statistically significant changes in the number of ED visits. Phillips et al. (2006) provided case management to high-risk patients manually selected by physicians and found a statistically significant *increase* in the number of ED visits afterward. These mixed results strongly suggest that case management might be ineffective or even harmful when provided to the wrong subpopulation. Unfortunately, if the treatment is potentially harmful, scoring rules can perform arbitrarily badly.

Remark 1 (Scoring Rules Can be Worse Than not Targeting). Index patients as in Theorem 1, and suppose that $\delta_c/\hat{\delta}_c = \underline{\delta} < 0$ for all $c \in \{1, \cdots, K\}$ and $\sum_{c\in B^*} r_c\delta_c > 0$. Such a scenario might occur if the treatment only benefitted a small subgroup of the population but potentially harmed others, and scoring rules cannot perfectly determine which is which. Then, $r\hat{\delta}$-scoring has intervention effectiveness $\underline{\delta}\sum_{c=1}^{K} r_c\hat{\delta}_c < 0$, which, in an absolute sense, is worse than not providing treatment to anyone. In terms of relative performance, the performance can be arbitrarily bad when the treatment is only marginally effective since

$$\frac{\underline{\delta}\sum_{c=1}^{K} r_c\hat{\delta}_c}{\sum_{c\in B^*} r_c\delta_c} \to -\infty \text{ as } \sum_{c\in B^*} r_c\delta_c \to 0. \tag{3}$$

In terms of the performance difference, this may be as large as

$$\overline{\delta}\sum_{c=K+1}^{2K} r_c\hat{\delta}_c - \underline{\delta}\sum_{c=1}^{K} r_c\hat{\delta}_c, \tag{4}$$

if $B^* = \{K+1, \ldots, 2K\}$ and $\delta_c/\hat{\delta}_c = \overline{\delta} > 0$ for all $c \in B^*$. Such a scenario might occur if there do exist high-reward patients who would benefit from the treatment, but the particular scoring rule does not identify them.

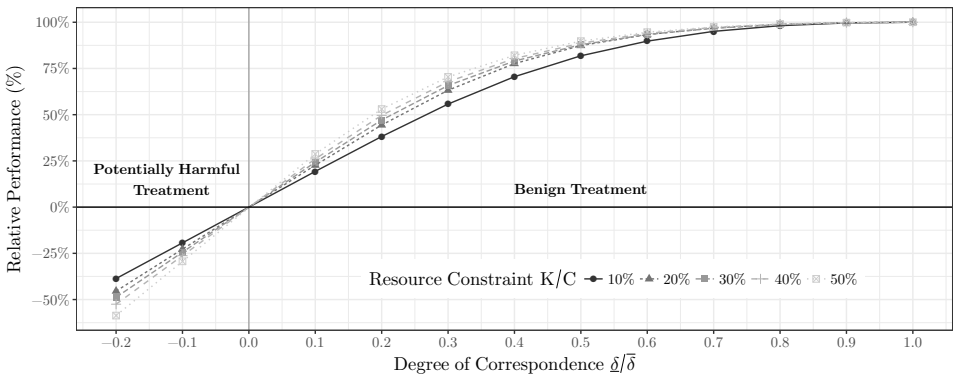
3.3. Case Study: Performance of Scoring Rules

Using patient data from our partner hospital, we apply Theorem 1 and Remark 1 for reward scoring ($\hat{\delta}_c = 1$) in Figure 1. Since $\hat{\delta}_c = 1$, $\underline{\delta}/\overline{\delta}$ measures the degree of heterogeneity in causal effects. For example, $\underline{\delta}/\overline{\delta} = 0.1$ means that the causal effects differ by a factor of 10, i.e., if there exist patients for whom case management reduces the number of ED visits by 1, there also exist patients for whom case management reduces the number of ED visits by 10.

When case management always reduces ED visits (i.e., benign treatment), reward scoring performs well and improves as the resource constraint is relaxed (K/C increases). For example, when

$\underline{\delta}/\overline{\delta} = 0.6$, reward scoring obtains approximately 90% of the full-information optimal value. As predicted by Corollary 1, the performance improves as the value of $\underline{\delta}/\overline{\delta}$ approaches 1. The particularly strong performance of reward scoring in this example is due to a small group of patients in our data who incur a very large charge per ED visit, i.e., r_c has a long tail. The 99^{th} percentile of r_c is 2.8 times the 90^{th} percentile value. Consequently, k^* is generally large.

Figure 1 Worst-Case Relative Performance of Reward Scoring (r-Scoring) for Case Management in Our Partner Hospital



Note. When the treatment is benign, we plot the worst-case relative performance bound (2) provided in Theorem 1. When the treatment is potentially harmful, the worst-case relative performance is $-\infty$, as mentioned in Remark 1. Thus, we plot $\underline{\delta}\sum_{c=1}^{K} r_c/\overline{\delta}\sum_{c=K+1}^{2K} r_c$ for comparison.

To the contrary, when case management may be ineffective or increase ED visits (i.e., potentially harmful treatment), reward scoring performs quite badly even when case management could increase the number of ED visits to a smaller degree compared to what it can reduce (e.g., $\underline{\delta}/\overline{\delta} = -0.1$). This poor worst-case performance is also caused by the long tail of r_c. Intuitively, when there exists a small proportion of patients with very high marginal rewards, targeting these patients is "risky." If those patients benefit from treatment, the overall effectiveness will be high, but if they respond negatively, the overall effectiveness will be low. Similar behavior is seen for outcome scoring in Figure EC.2 in the e-companion. Both figures highlight the fact that when the particular score has a long tail for a fixed degree of correspondence, scoring rules will be very sensitive to whether or not the treatment is potentially harmful.

4. Robust Targeting

The worst-case performance of scoring rules depends strongly on whether the treatment is potentially harmful. We next introduce our robust approach, which is less sensitive to this distinction.

4.1. Similar Patients Respond Similarly

The key idea of our approach stems from the simple intuition that *patients with similar pre-treatment covariates should respond similarly to treatment.* To formalize this idea, we first define the Conditional Average Treatment Effect (CATE) and re-express Eq. (1).

To avoid writing long summations in what follows, we define the random variables $\tilde{c}$ to be a randomly chosen patient from the candidate population. Thus, $(\mathbf{x}_{\tilde{c}}, \delta_{\tilde{c}})$ denote the pre-treatment covariates and causal effect for this randomly chosen patient. Define $\tilde{s}$ and $(\mathbf{x}^{\tilde{s}}, \delta^{\tilde{s}})$ similarly for the study population. We stress that $\tilde{c}$ and $\tilde{s}$ are only defined to simplify the notation; the values $(\mathbf{x}_c, \delta_c)$ $c = 1 \ldots, C$ and $(\mathbf{x}^s, \delta^s)$ $s = 1, \ldots, S$ are fixed, unknown constants, i.e., non-random. With this notation, the SATE in the study population can now be concisely expressed as $\frac{1}{S}\sum_{s=1}^{S} \delta^s = \mathbb{E}[\delta^{\tilde{s}}]$.

Let the study population CATE for a patient with pre-treatment covariates $\mathbf{x} \in \mathcal{X}$ be

$$\mathbb{E}[\delta^{\tilde{s}} | \mathbf{x}^{\tilde{s}} = \mathbf{x}] = \frac{1}{|\{s \mid \mathbf{x}^s = \mathbf{x}\}|} \sum_{s:\mathbf{x}^s = \mathbf{x}} \delta^s.$$

The study population CATE is a function of $\mathbf{x}$. Define the candidate population CATE, i.e., $\mathbb{E}[\delta_{\tilde{c}} | \mathbf{x}_{\tilde{c}} = \mathbf{x}]$, similarly. Intuitively, CATE represents the average causal effect across all patients in the given population with a particular value of covariate.

Using CATE, we can re-express the objective of Eq. (1). Recall, r_c is a component of $\mathbf{x}_c$, so that $r_{\tilde{c}}$ is $\mathbf{x}_{\tilde{c}}$ measurable. Suppose that $z_1, \ldots z_C$ represent a targeting policy where z_c depends only on $\mathbf{x}_c$, i.e, $z_{\tilde{c}}$ is $\mathbf{x}_{\tilde{c}}$ measurable. Then, the objective of Eq. (1) for this policy is

$$\sum_{c=1}^{C} z_c r_c \delta_c \;=\; C \cdot \mathbb{E}\left[z_{\tilde{c}} r_{\tilde{c}} \delta_{\tilde{c}}\right] \;=\; C \cdot \mathbb{E}\Big[z_{\tilde{c}} r_{\tilde{c}} \mathbb{E}\left[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}}\right]\Big] \;=\; \sum_{c=1}^{C} z_c r_c \mathbb{E}\left[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c\right], \tag{5}$$

where the first and last equalities follow from the definition of $\tilde{c}$, and the middle equality uses that $r_{\tilde{c}}$ is $\mathbf{x}_{\tilde{c}}$ measurable. Thus, the objective of Eq. (1) is equivalent to the objective of:

$$\max_{\mathbf{z} \in \mathcal{Z}} \sum_{c=1}^{C} z_c r_c \mathbb{E}[\delta_{\tilde{c}} | \mathbf{x}_{\tilde{c}} = \mathbf{x}_c]. \tag{6}$$

Replacing the objective of Eq. (1) with the objective of Eq. (6) is conceptually appealing. Recall, δ_c are fundamentally unobservable since we cannot observe *both* $y_c(0)$ and $y_c(1)$. By contrast, $\mathbb{E}\left[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}\right]$ *is* estimable given a large enough RCT in the candidate population. This is perhaps why most personalization schemes focus on Eq. (6) directly (see, e.g., Kallus 2017, Athey and Wager 2017). Moreover, via a similar argument, we can rewrite the confidence interval for the study SATE as a constraint on the study CATE, i.e., $\mathbb{E}[\delta^{\tilde{s}}] \in [\underline{I}, \overline{I}] \iff \mathbb{E}\left[\mathbb{E}[\delta^{\tilde{s}} \mid \mathbf{x}^{\tilde{s}}]\right] \in [\underline{I}, \overline{I}]$, and focus on CATEs exclusively.

The challenge with Eq. (6) in our setting, however, is that we do not know the candidate population CATE; our data on effectiveness is from the study population. We must assume some "link" between the candidate CATE and the study CATE in order to leverage the study evidence.

To that end, fix any norm $\|\cdot\|_{\mathsf{link}}$ on $\mathbb{R}^C$ and define the constant κ by

$$\kappa \equiv \left\| \left(\mathbb{E}[\delta^{\tilde{s}} \mid \mathbf{x}^{\tilde{s}} = \mathbf{x}_c] - \mathbb{E}[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c] \right)_{c=1}^{C} \right\|_{\mathsf{link}}. \tag{7}$$

In words, κ measures an aggregate distance between the study CATE and candidate CATE on the candidate population. Importantly, κ makes rigorous the idea that similar patients in both populations should respond similarly to treatment. For example, when $\kappa = 0$, the CATEs are identical in both populations, and the expected treatment effectiveness of a patient given her pre-treatment covariates does not depend on her population. Positive, but small, κ bounds the difference in effect between the populations.[3]

4.2. A First Robust Model

Since the candidate CATE is unknown, one approach might be to model this CATE as a random function, say $\tilde{\Psi}_{\mathsf{C}}(\cdot)$, and then seek to solve $\max_{\mathbf{z}\in\mathcal{Z}} \sum_{c=1}^{C} z_c r_c \mathbb{E}\left[\tilde{\Psi}_{\mathsf{C}}(\mathbf{x}_c)\right]$. Unfortunately, the data at hand do not contain information about the precise *structure* of the candidate CATE. Consequently, defining the probability distribution of $\tilde{\Psi}_{\mathsf{C}}(\cdot)$ would require strong a priori assumptions on this structure that may not be easily validated.

Instead, we adopt a robust optimization perspective. Specifically, we maximize the worst-case intervention effectiveness over possible values for the candidate CATE that are consistent with the study findings:

$$\max_{\mathbf{z}\in\mathcal{Z}} \min_{\Psi_{\mathsf{C}}(\cdot)\in\mathcal{U}} \sum_{c=1}^{C} z_c r_c \Psi_{\mathsf{C}}(\mathbf{x}_c), \tag{8}$$

where $\Psi_{\mathsf{C}}(\cdot)$ approximates the candidate CATE. Given our study evidence, a first choice for the uncertainty set $\mathcal{U}$ might be

$$\mathcal{U}_{\hat{\kappa}} = \left\{ \Psi_{\mathsf{C}} : \mathcal{X} \mapsto \mathbb{R} \;\middle|\; \exists \Psi^{\mathsf{S}} : \mathcal{X} \mapsto \mathbb{R} \text{ s.t. } \underline{I} \le \mathbb{E}[\Psi^{\mathsf{S}}(\mathbf{x}^{\tilde{s}})] \le \overline{I}, \;\; \left\| \left(\Psi^{\mathsf{S}}(\mathbf{x}_c) - \Psi_{\mathsf{C}}(\mathbf{x}_c)\right)_{c=1}^{C} \right\|_{\mathsf{link}} \le \hat{\kappa} \right\}, \tag{9}$$

where $\Psi^{\mathsf{S}}(\cdot)$ approximates the study CATE, and the user-defined parameter $\hat{\kappa}$ approximates κ.

Unfortunately, this uncertainty set does not yield practically implementable solutions. Specifically, for any fixed $\mathbf{z} \in \{0,1\}^C$, $\mathbf{z} \neq \mathbf{0}$ and any $\mathbf{x} \in \mathcal{X}$, let $q_{\mathbf{z}}(\mathbf{x}) \equiv \sum_{c=1}^{C} z_c r_c \mathbb{I}(\mathbf{x}_c = \mathbf{x}) / \sum_{c=1}^{C} z_c r_c$ denote the reward-weighted covariate distribution of the targeted patients from the candidate population. Define the set

$$\mathcal{Z}^* \equiv \left\{ \mathbf{z} \in \{0,1\}^C \;\middle|\; \sum_{c=1}^{C} z_c \le K, \;\; q_{\mathbf{z}}(\mathbf{x}) = \mathbb{P}(\mathbf{x}^{\tilde{s}} = \mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \right\}.$$

THEOREM 2 **(Trivial Solutions for Unbounded CATEs)**. *For any $\hat{\kappa} \ge 0$, either $\mathbf{0}$ is an optimal solution to (8) with $\mathcal{U}_{\hat{\kappa}}$ or every optimal solution is contained in $\mathcal{Z}^*$.*

[3] The assumption that $\kappa = 0$ and the CATEs are identical is common in statistical techniques that generalize a causal effect from one population to another (e.g., Cole and Stuart 2010, Stuart et al. 2011, Hartman et al. 2015). Our proposed procedure will not depend on the true value of κ, although its performance will (see Corollary 5). Consequently, we will formulate our model in the general case when $\kappa \ge 0$ and the study and candidate CATE may differ, but on first reading, the reader may assume $\kappa = 0$ without much loss of generality.

Theorem 2 asserts that under uncertainty set (9), the worst-case optimal targeting either chooses no patients or matches the study distribution of covariates exactly. Verifying that a solution matches the study distribution of covariates exactly, however, requires access to the full distribution of $\mathbf{x}^{\tilde{s}}$, i.e., the raw study data, making it impossible to compute such a solution.

4.3. Incorporating Description Functions

The fundamental issue is that $\mathcal{U}_{\hat{\kappa}}$ in (9) is "too large" and contains many pathological pairs $\Psi_{\mathsf{C}}, \Psi^{\mathsf{S}}$. Ideally, we would prefer to restrict $\mathcal{U}_{\hat{\kappa}}$ to a suitably well-behaved, nonparametric class of functions, such as those belonging to a kernel space (Kallus 2017). However, for an arbitrary nonparametric specification of the study CATE, verifying $\mathbb{E}[\delta^{\tilde{s}}] \in [\underline{I}, \overline{I}]$ may require the full distribution of $\mathbf{x}^{\tilde{s}}$. Consequently, we restrict $\mathcal{U}_{\hat{\kappa}}$ to a particular nonparametric class of functions for which we can easily verify this condition.

To this end, consider projecting the study CATE onto the affine space spanned by the description functions $\{1, \phi_1(\cdot), \ldots, \phi_G(\cdot)\}$. Then, for any $\mathbf{x} \in \mathcal{X}$, we can write $\mathbb{E}[\delta^{\tilde{s}} \mid \mathbf{x}^{\tilde{s}} = \mathbf{x}] = \beta_0^* + \sum_{g=1}^G \beta_g^* \phi_g(\mathbf{x}) + \epsilon^*(\mathbf{x})$, where

$$(\beta_0^*, \boldsymbol{\beta}^*) \in \arg\min_{\beta_0, \boldsymbol{\beta}} \quad \left\| \left(\mathbb{E}[\delta^{\tilde{s}} \mid \mathbf{x}^{\tilde{s}} = \mathbf{x}^s] - \beta_0 - \sum_{g=1}^G \beta_g \phi_g(\mathbf{x}^s) \right)_{s=1}^S \right\|_2^2. \tag{10}$$

Here $\|\cdot\|_2$ is the ordinary ℓ_2-norm. Consequently, by construction, $\mathbb{E}[\epsilon^*(\mathbf{x}^{\tilde{s}})] = 0$.

By itself, this decomposition does not restrict the set of CATEs under consideration; any CATE can be projected onto this affine subspace. This is why we describe our specification as nonparametric. Nonetheless, because $\mathbb{E}[\epsilon^*(\mathbf{x}^{\tilde{s}})] = 0$, we have that $\mathbb{E}[\Psi^{\mathsf{S}}(\mathbf{x}^{\tilde{s}})] = \beta_0^* + \sum_{g=1}^G \beta_g^* \mathbb{E}[\phi_g(\mathbf{x}^{\tilde{s}})] = \beta_0^* + \sum_{g=1}^G \beta_g^* \mu_g$. Thus, despite the non-parametric specification, verifying the study SATE agrees with the evidence only requires knowing $\beta_0^*, \boldsymbol{\beta}^*$, i.e., $\mathbb{E}[\delta^{\tilde{s}}] \in [\underline{I}, \overline{I}] \iff \beta_0^* + \sum_{g=1}^G \beta_g^* \mu_g \in [\underline{I}, \overline{I}]$.

Motivated by this decomposition, our new uncertainty set is formed by restricting the size of the coefficients and residual in this decomposition. Specifically, let $\|\cdot\|$ be a norm on $\mathbb{R}^G$ and $\|\cdot\|_{\mathsf{res}}$ be a norm on $\mathbb{R}^C$. The main uncertainty set of our robust model is then

$$\begin{aligned} \mathcal{U}_{\hat{\boldsymbol{\Gamma}}, \hat{\kappa}} = \Bigg\{ \Psi_{\mathsf{C}}(\cdot) : \mathcal{X} \mapsto \mathbb{R} \;\Bigg|\; & \exists \epsilon(\cdot) : \mathcal{X} \mapsto \mathbb{R},\ \beta_0 \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^G,\ \text{s.t.}\ \ \Psi^{\mathsf{S}}(\mathbf{x}) = \beta_0 + \sum_{g=1}^G \beta_g \phi_g(\mathbf{x}) + \epsilon(\mathbf{x}), \\ & \underline{I} \le \beta_0 + \sum_{g=1}^G \beta_g \mu_g \le \overline{I},\ \ \left\| \left(\Psi^{\mathsf{S}}(\mathbf{x}_c) - \Psi_{\mathsf{C}}(\mathbf{x}_c) \right)_{c=1}^C \right\|_{\mathsf{link}} \le \hat{\kappa},\ \ \|\boldsymbol{\beta}\| \le \hat{\Gamma}_1,\ \ \left\| (\epsilon(\mathbf{x}_c))_{c=1}^C \right\|_{\mathsf{res}} \le \hat{\Gamma}_2, \Bigg\}. \end{aligned} \tag{11}$$

In words, the first equality of our uncertainty set decomposes $\Psi^{\mathsf{S}}(\cdot)$ into its projection onto the affine space of description functions and a residual. Any function can be decomposed in this way, so this equality does not limit the class of study CATEs under consideration. The second pair

of inequalities model the study evidence. The third inequality bounds the distance between the CATEs as in Eq. (9). The last two inequalities depend on the user-defined parameters $\hat{\Gamma}_1, \hat{\Gamma}_2$ and restrict the set of possible CATEs beyond Eq. (9). Also note that this uncertainty set is always non-empty for any nonnegative $(\hat{\kappa}, \hat{\Gamma}_1, \hat{\Gamma}_2)$. One can verify that $\Psi_{\mathsf{C}}(\mathbf{x}) = \frac{\underline{I}+\overline{I}}{2}$ for all $\mathbf{x} \in \mathcal{X}$ is a member of $\mathcal{U}_{\hat{\mathbf{\Gamma}},\hat{\kappa}}$ by letting $\Psi_{\mathsf{C}}(\cdot) = \Psi^{\mathsf{S}}(\cdot)$, $\beta_0 = \frac{\underline{I}+\overline{I}}{2}$, $\boldsymbol{\beta} = \mathbf{0}$ and $\epsilon(\cdot) = 0$. In other words, for any choices of the model parameters, our uncertainty set includes a "nominal" case where there is no heterogeneity in causal effects, and the SATE of the study and targeted population are the same.

To build some intuition for these last two constraints, consider the idealized special case where we 1) choose the norm $\|\cdot\|$ such that $\|\boldsymbol{\beta}\|^2 \equiv \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}$ with $\Sigma_{gg'} \equiv \mathbb{E}[(\phi_g(\mathbf{x}^{\tilde{s}}) - \mu_g)(\phi_{g'}(\mathbf{x}^{\tilde{s}}) - \mu_{g'})]$ for all $g, g' = 1, \ldots, G$ and 2) choose the norm $\|\cdot\|_{\mathsf{res}}$ such that $\left\|(\epsilon(\mathbf{x}_c))_{c=1}^C\right\|_{\mathsf{res}} = \left\|(\epsilon(\mathbf{x}^s))_{s=1}^S\right\|_2$.[4] Then, the constraint $\|\boldsymbol{\beta}\| \leq \hat{\Gamma}_1$ is equivalent to $\mathrm{Var}(\beta_0 + \sum_{g=1}^G \beta_g \phi_g(\mathbf{x}^{\tilde{s}})) \leq \hat{\Gamma}_1^2$, and the constraint on $\epsilon(\cdot)$ bounds the residual variance in the regression Eq. (10). In other words, $\hat{\Gamma}_1$ controls the amount of variability of $\Psi(\mathbf{x}^{\tilde{s}})$ explained by the description functions, while $\hat{\Gamma}_2$ controls the residual variability.

With these choices of norm, the sum $\hat{\Gamma}_1 + \hat{\Gamma}_2$ is the total variance of $\Psi(\mathbf{x}^{\tilde{s}})$ and describes the heterogeneity in the study CATE. As this sum tends to zero, $\Psi(\mathbf{x}^{\tilde{s}})$ tends to a constant, i.e., the study-CATE is homogenous. At the same time, the ratio $\frac{\hat{\Gamma}_1}{\hat{\Gamma}_1+\hat{\Gamma}_2}$ describes the ability of these description functions to capture this hetereogeneity, much like the R^2 of a linear regression. If this ratio is large, these description functions capture most of the heterogeneity, and given $\phi_g(\mathbf{x}^s)$ we can predict the CATE of patient s well. If this ratio is small, these description functions are uninformative, and we cannot predict the CATE of patient s well from these values.

That said, we stress these choices for norms are idealized, not prescriptive. Specifying the norms in this manner would require detailed information on the distribution of covariates in the study population, which is unavailable. Nonetheless, we will leverage this intuition to motivate specific, practical choices of the norm in special cases in what follows.

4.4. Robust Counterparts

Using standard techniques, we can compute a robust counterpart. Let $\|\cdot\|^*$, $\|\cdot\|_{\mathsf{link}}^*$ and $\|\cdot\|_{\mathsf{res}}^*$ be dual norms to $\|\cdot\|$, $\|\cdot\|_{\mathsf{link}}$ and $\|\cdot\|_{\mathsf{res}}$, respectively.

THEOREM 3 **(General Robust Counterpart)**. *The robust targeting problem* (8) *with uncertainty set* (11) *is equivalent to*

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad \underline{I} \sum_{c=1}^C z_c r_c - \hat{\Gamma}_1 \left\| \left(\sum_{c=1}^C z_c r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^G \right\|^* - \hat{\Gamma}_2 \left\| (z_c r_c)_{c=1}^C \right\|_{\mathsf{res}}^* - \hat{\kappa} \left\| (z_c r_c)_{c=1}^C \right\|_{\mathsf{link}}^*. \quad (12)$$

[4] This norm can always be specified in this manner whenever the random variables $\mathbf{x}^{\tilde{s}}$ and $\mathbf{x}_{\tilde{c}}$ are mutually absolutely continuous. In this case, simply take $\left\|(\epsilon(\mathbf{x}_c))_{c=1}^C\right\|_{\mathsf{res}}^2 \equiv \sum_{c=1}^C \epsilon(\mathbf{x}_c)^2 \frac{\mathbb{P}(\mathbf{x}^{\tilde{s}}=\mathbf{x}_c)}{\mathbb{P}(\mathbf{x}_{\tilde{c}}=\mathbf{x}_c)}$. One can then check directly that $\left\|(\epsilon(\mathbf{x}_c))_{c=1}^C\right\|_{\mathsf{res}} = \left\|(\epsilon(\mathbf{x}^s))_{s=1}^S\right\|_2$.

Remark 2 (Computational Complexity). From a theoretical point of view, Problem (12) is *NP-Complete*, even when $G = 1$, $\hat{\kappa} = \hat{\Gamma}_2 = 0$, $\phi(\mathbf{x})$ takes binary values and $\|\cdot\|^*$ is an ℓ_p-norm (Theorem EC.1, Appendix EC.1). From a practical point of view, when the norms correspond to (weighted) ℓ_1 or ℓ_∞-norms, Problem (12) is a mixed-binary linear program, and when the norms corresponds to (weighted) ℓ_2-norms, Problem (12) is a mixed-binary second-order cone problem. Although theoretically difficult, moderately sized mixed-binary linear and mixed-binary second-order cone problems (such as those we study in this paper) can be solved efficiently using off-the-shelf software on a personal computer in minutes.

Problem (12) provides insight into the structure of an optimal targeting:

No Dependence on $\overline{I}$. The robust counterpart does not depend on $\overline{I}$, or, equivalently, the width of the confidence interval $\overline{I} - \underline{I}$. This lack of dependence is a unique feature of our causal inference setting that distinguishes it from more traditional data-driven robust optimization settings.

Specifically, in typical data-driven settings where one directly observes data on the relevant uncertainties, the width of the of the confidence interval roughly corresponds to the precision of the estimates of those uncertainties (see, e.g., Bertsimas et al. 2017). With more data, this interval shrinks, and the robust counterpart "converges" to a nominal (full-information) problem.

In our setting, we do not directly observe data on the relevant uncertainty, i.e., the candidate CATE. The width of the confidence interval $[\underline{I}, \overline{I}]$ does *not* correspond to the precision of the relevant uncertain parameters, i.e., the candidate CATE, but rather to the precision of the estimate for the study SATE. The precision of this estimator does not affect our targeting. What matters in the targeting problem is the level of the study SATE and the variability of study CATE. Intuitively, an RCT with an extremely large sample size could drive the width of the confidence interval to zero, but that would not imply that the study CATE had low variability. There would *still* be uncertainty in the form of the heterogenous effect in candidate population.

Avoiding high-reward patients if $\hat{\kappa}$ or $\hat{\Gamma}_2$ is large. As $\hat{\kappa} \to \infty$ with $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$ fixed, the last term in Eq. (12) grows. Consequently, an optimal solution selects fewer and fewer patients with very high rewards, until ultimately no patients are targeted. We claim this behavior is intuitive. Recall that $\hat{\kappa}$ proxies κ (see Eq. (7)). If a decision-maker believed κ were quite large (and hence specified $\hat{\kappa}$ to be large), she also believes that the study CATE is not representative of the candidate CATE. Said another way, she believes there is only weak evidence in the study, itself, to guarantee that targeting in the candidate population will be effective. Consequently, an optimal risk-averse targeting should select relatively few patients, and avoid patients with very high rewards. Indeed, recall from our discussion of the pitfalls of reward scoring around Fig. 1 that high-reward patients are "risky." If these patients react adversely to treatment, they have very negative effectiveness.

The robust model guards against the pitfalls of reward-scoring when the study-evidence is weak. If the evidence is weak enough, it recommends not targeting.

A similar behavior holds as $\hat{\Gamma}_2 \to \infty$ with $\hat{\kappa}$ and $\hat{\Gamma}_1$ fixed. Recall that $\hat{\Gamma}_2$ proxies the residual error in Eq. (10). If a decision-maker believed Γ_2 were large (and hence specified a large $\hat{\Gamma}_2$), then even if she believed $\kappa = 0$, i.e., that study CATE and candidate CATE were identical, she should select relatively few patients and avoid patients with high rewards. Indeed, asserting $\kappa = 0$ is tantamount to saying the way causal effects depend on covariates $\mathbf{x}$ is the same in both populations. However, asserting that Γ_2 is large implies that this dependence relies on information in $\mathbf{x}$ not captured by the values $\phi_1(\mathbf{x}), \ldots, \phi_G(\mathbf{x})$. Consequently, there is still only weak evidence in the study, itself, to guarantee that targeting in the candidate population will be effective. Note the dependence of Problem (12) on $\hat{\Gamma}_1$ is more subtle and discussed in detail in the next subsection.

Although Theorem 3 is stated in full-generality, it depends on three user-defined parameters $\hat{\kappa}$, $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$, and three user-defined norms $\|\cdot\|$, $\|\cdot\|_{\mathsf{link}}$ and $\|\cdot\|_{\mathsf{res}}$. Practically, it is not clear that the data at hand, i.e., the summary statistics and the study-evidence, support this detailed a specification. We might prefer a simpler model with fewer parameters to specify.

Consequently, in what follows, we propose choosing $\|\cdot\|_{\mathsf{res}}$ and $\|\cdot\|_{\mathsf{link}}$ to be ℓ_∞-norms. The resulting counterpart has a simple form, with only one effective user-defined parameter and norm.

Corollary 2 **(Simplified Robust Counterpart)**. *Suppose both $\|\cdot\|_{\mathsf{res}}$ and $\|\cdot\|_{\mathsf{link}}$ are taken to be ℓ_∞-norms. Let $\mathbf{z}^*$ be an optimal solution to problem* (8) *with uncertainty set* (11)*. Then,*

1. *If $\underline{I} - \hat{\Gamma}_2 - \hat{\kappa} \le 0$, $\mathbf{z}^* = \mathbf{0}$.*
2. *If $\underline{I} - \hat{\Gamma}_2 - \hat{\kappa} > 0$, $\mathbf{z}^*$ is also an optimal solution to*

$$\max_{\mathbf{z}\in\mathcal{Z}} \sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \left\| \left(\sum_{c=1}^{C} z_c r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*, \tag{13}$$

and the optimal value of problem (8) *is $(\underline{I} - \hat{\Gamma}_2 - \hat{\kappa})$ time the optimal value of problem* (13)*.*

We argue that problem (13) represents a good *practical* modeling compromise. The simplified structure still captures many of the qualitative features of problem (12), e.g., for sufficiently large $\hat{\kappa}$ or $\hat{\Gamma}_2$, we should target no one. More importantly, finding an optimal solution only requires specifying one user-defined parameter, i.e., the ratio $\frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}}$ and one user-defined norm, i.e., $\|\cdot\|$.

This ratio further admits a simple interpretation as an "adjusted" coefficient of variation (CV). Specifically, in the special case when $\hat{\kappa} = \hat{\Gamma}_2 = 0$, then, under our earlier idealized choice of norm $\|\boldsymbol{\beta}\|^2 \equiv \mathrm{Var}\left(\sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}^{\tilde{s}})\right)$, this ratio upper bounds the coefficient of variation of our approximate study CATE $\Psi^{\mathsf{S}}(\mathbf{x}^{\tilde{s}})$. Equivalently, since $\hat{\kappa} = 0$, it also upper bounds the coefficient of variation on our approximate candidate CATE on the study population $\Psi_{\mathsf{C}}(\mathbf{x}^{\tilde{s}})$. When $\kappa > 0$ or $\hat{\Gamma}_2 > 0$, we adjust this coefficient of variation by reducing our estimate of the mean effectiveness due to differences between the study and candidate populations, i.e., reducing $\underline{I}$ to $\underline{I} - \Gamma_2 - \hat{\kappa}$.

4.5. Covariate Matching as Regularization

Problem (13) also facilitates a new connection between our approach and covariate matching in statistics. Intuitively, we might expect that the larger the difference between the covariates of the study population and targeted patients is, the less reliable the SATE of the study population is as an estimate of causal effects in the targeted patients. Thus, we should avoid such targetings.

To make this intuition precise, note when $\underline{I} - \hat{\Gamma}_2 - \hat{\kappa} \geq 0$, we can rewrite Eq. (13) as

$$\sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \cdot \left\| \left(\sum_{c=1}^{C} w_c \phi_g(\mathbf{x}_c) - \mu_g \right)_{g=1}^{G} \right\|_* \cdot \sum_{c=1}^{C} z_c r_c, \quad \text{where } w_c \equiv \frac{z_c r_c}{\sum_{c=1}^{C} z_c r_c}. \tag{14}$$

Thus, the objective Problem (13) approximates the intervention effectiveness of a candidate targeting $\mathbf{z}$ by its total reward *minus* a penalty that depends on the distance between the summary statistics $\boldsymbol{\mu}$ evaluated on the study population and a reward-weighted average of these statistics $(\sum_{c=1}^{C} w_c \phi_g(\mathbf{x}_c))_{g=1}^{G}$ evaluated on the targeted patients from the candidate population. Our adjusted CV $\frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}}$ controls the trade-off between these two objectives. For small adjusted CV, solutions to (13) will target high-reward patients regardless of their covariates. When the adjusted CV is 0, problem (13) reduces to reward scoring. As the adjusted CV increases, solutions to (13) will match the reward-weighted average summary statistics in the study population more closely. Similar behavior holds for the general problem (12).

This "covariate matching as regularization" interpretation of our model provides a natural intuition that connects with the literature on matching in design of experiments. Unlike traditional schemes for matching, however, our approach incorporates the rewards r_c both in the objective and in the particular structure of the penalty. We next show that in special cases with appropriately chosen norms $\|\cdot\|$, we recover common matching procedures in the regularizer.

For any positive definite matrix $\mathbf{A}$, let $\|\mathbf{t}\|_{\mathbf{A}} \equiv \sqrt{\mathbf{t}^T \mathbf{A} \mathbf{t}}$. The corresponding dual norm is $\|\cdot\|_{\mathbf{A}^{-1}}$.

Corollary 3 (**χ^2-Matching under Partition Description Functions**). *Suppose there exists a partition $\mathcal{X} = \bigcup_{g=1}^{G+1} \mathcal{X}_g$, and $\phi_g(\mathbf{x}) = \mathbb{I}(\mathbf{x} \in \mathcal{X}_g)$ are our description functions with statistics μ_g for all $g = 1, \ldots, G$. Define $\|\cdot\|$ in (11) by $\|\cdot\| \equiv \|\cdot\|_{\boldsymbol{\Sigma}}$, where $\boldsymbol{\Sigma} \equiv diag(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^T \in \mathbb{R}^{G \times G}$. Then, Eq.* (13) *with this uncertainty set is equivalent to*

$$\sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \cdot \sqrt{\sum_{g=1}^{G+1} \frac{(q_{\mathbf{z},g} - \mu_g)^2}{\mu_g}} \cdot \sum_{c=1}^{C} z_c r_c, \tag{15}$$

where $q_{\mathbf{z},g} \equiv \sum_{c=1}^{C} z_c r_c \mathbb{I}(\mathbf{x}_c \in \mathcal{X}_g) / \sum_{c=1}^{C} z_c r_c$ is the (reward-weighted) proportion of type g patients in the targeted population and $\mu_{G+1} \equiv 1 - \sum_{g=1}^{G} \mu_g$.

The penalty term of (15) is the χ^2-distance between $(\mu_g)_{g=1}^{G+1}$ and $(q_{\mathbf{z},g})_{g=1}^{G+1}$. The χ^2-distance metric is commonly used for matching with partitioned covariates (Imbens and Rubin 2015). It arises naturally as a regularizer in our method through the appropriate choice of uncertainty set.

REMARK 3. When $|\mathcal{X}|$ is finite and the partition consists of singletons, *every* CATE $\Psi^{\mathsf{S}}(\mathbf{x})$ can be written in the form $\Psi^{\mathsf{S}}(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \mathbb{I}(\mathbf{x} \in \mathcal{X}_g)$ for some $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{G+1}$. Consequently, one can take $\Gamma_2 = \hat{\Gamma}_2 = 0$ without loss of generality. This is the canonical setting for Corollary 3.

COROLLARY 4 **(Mean Matching under Linear Description Functions)**. *Suppose* $\mathcal{X} \in \mathbb{R}^G$*, and* $\phi_g(\mathbf{x}) = x_g$ *are our description functions with statistics* μ_g *for all* $g = 1, \ldots, G$*. For any positive definite matrix* $\mathbf{V} \in \mathbb{R}^{G \times G}$*, consider uncertainty set* (11) *with a weighted* ℓ_2*-norm* $\|\cdot\|_V$*. Then, Problem* (13) *is equivalent to*

$$\max_{\mathbf{z} \in \mathcal{Z}} \quad \sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \cdot \left\| \left(\sum_{c=1}^{C} w_c \mathbf{x}_c - \mu_g \right)_{g=1}^{G} \right\|_{\mathbf{V}^{-1}} \cdot \sum_{c=1}^{C} z_c r_c, \tag{16}$$

where $w_c \equiv \frac{z_c r_c}{\sum_{c=1}^{C} z_c r_c}$.

The penalty term of Eq. (16) is a (weighted) distance between the means of the covariates in the study population and in the target population. These types of distances between means are frequently used to assess the quality of covariate matching (Imbens and Rubin 2015, pg. 410). The choice of $\mathbf{V}$ controls the weighting. A common choice is to take $\mathbf{V}$ to be $\boldsymbol{\Sigma}$, the covariance matrix of $\mathbf{x}^{\tilde{s}}$, which recovers so-called Mahalanobis matching (Imbens and Rubin 2015, pg. 411). Another common choice is to take $\mathbf{V}$ to be $\text{diag}(\boldsymbol{\Sigma})$. This choice recovers the Euclidean metric (Imbens and Rubin 2015, pg. 411) or so-called mean-matching penalty (Kallus 2016).

In summary, the interpretation of our method as regularizing via covariate matching highlights that the role of the norm $\|\cdot\|^*$ is primarily to establish a metric between the distribution of covariates in the study population and the (reward-weighted) distribution of covariates in the targeted group. Indeed, any choice of norm enjoys this interpretation. Thus, although it is certainly mathematically elegant to let $\|\cdot\|$ to be $\|\cdot\|_{\boldsymbol{\Sigma}}$, when $\boldsymbol{\Sigma}$ is known, other reasonable norms should still yield good performance. Indeed, we use $\text{diag}(\boldsymbol{\Sigma})$ to specify the norm in our case-study, because the full covariance matrix of covariates is not reported in Shumway et al. (2008).

In principle, one might ask if is possible to choose an uncertainty set to recover other covariate matching techniques. Appendix EC.2.4 describes a general construction. However, we consider this construction to be principally of theoretical, rather than practical, interest for two reasons: First, the above matching techniques (χ^2-matching, Mahalanobis Matching and mean-matching) are by far the most common in practice. Second, and more importantly, other covariate matching techniques typically require knowledge of the full-distribution of covariates, not simply knowledge of a few statistics. Since most studies do not report this full distribution, they cannot be used practically as a regularizer in our setting. We refer the readers to Appendix EC.2.4 for further details.

4.6. Performance Guarantee for the Robust Approach

Recall that for a sufficiently large radius, our uncertainty set contains the true candidate CATE. Using this observation, we bound the performance of our robust approach.

COROLLARY 5 **(Worst-Case Performance of Robust Targeting)**. *Let* $\Gamma_1, \Gamma_2, \kappa$ *be sufficiently large so that* $\mathbb{E}[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c] \in \mathcal{U}_{\mathbf{\Gamma},\kappa}$. *Let* $\mathbf{z}^{Rob}$ *be an optimizer of problem* (12) *for uncertainty set* $\mathcal{U}_{\hat{\mathbf{\Gamma}},\hat{\kappa}}$. *Let*

$$d_1 \equiv \left\| \left(\sum_{c=1}^{C} z_c^{Rob} r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*, \quad d_2 \equiv \left\| \left(z_c^{Rob} r_c \right)_{c=1}^{C} \right\|_{\mathsf{res}}^*, \quad d_3 \equiv \left\| \left(z_c^{Rob} r_c \right)_{c=1}^{C} \right\|_{\mathsf{link}}^*.$$

Then,

$$\sum_{c=1}^{C} z_c^{Rob} r_c \mathbb{E}[\delta_{\tilde{c}} | \mathbf{x}_{\tilde{c}} = \mathbf{x}_c] \;\geq\; \underline{I} \sum_{c=1}^{C} z_c^{Rob} r_c - \Gamma_1 d_1 - \Gamma_2 d_2 - \kappa d_3 \;\geq\; (\hat{\Gamma}_1 - \Gamma_1) d_1 + (\hat{\Gamma}_2 - \Gamma_2) d_2 + (\hat{\kappa} - \kappa) d_3.$$

Corollary 5 describes the performance of the robust model under misspecification of the parameters. The bound only depends on the unknown causal effect through the parameters $\Gamma_1, \Gamma_2, \kappa$. Corollary 5 guarantees that *if* we specify an uncertainty set large-enough, the robust strategy has non-negative effectiveness, i.e., it is not harmful. (A sufficient condition is that $\hat{\Gamma}_1 \geq \Gamma_1$, $\hat{\Gamma}_2 \geq \Gamma_2$ and $\hat{\kappa} \geq \kappa$.) This is structurally different from targeting rules, where the performance depends strongly on how well the rule matches the underlying causal effect if treatments may be harmful (Remark 1). Moreover, we stress that the bound only depends on the covariates differences between the study population and the targeted group, *not* the candidate population.

4.7. Selecting the Adj. CV parameter

Thus far, we have not discussed the choice of the adj. CV parameter $\frac{\hat{\Gamma}_1}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}}$. One approach might be to use *external* information or domain knowledge to 1) estimate the amount of explained heterogeneity in causal effects, 2) estimate the average causal effect in candidate population, and then 3) take their ratio. This approach requires external information or domain knowledge because the reported study-data itself does not contain information about these parameters.

We adopt a different viewpoint motivated by the satisficing literature (Simon 1955). Let

$$\mathbf{z}(\lambda) \in \arg\max_{\mathbf{z} \in \mathcal{Z}} \sum_{c=1}^{C} z_c r_c - \lambda \left\| \left(\sum_{c=1}^{C} z_c r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*.$$

Note that $\mathbf{z}(0)$ is the reward-scoring solution.

Then, given an acceptable revenue loss $0 < \alpha < 1$, we set λ to be the solution of

$$\max_{\lambda \geq 0} \quad \lambda \qquad \text{s.t.} \quad \sum_{c=1}^{C} r_c z_c(\lambda) \geq (1 - \alpha) \sum_{c=1}^{C} r_c z_c(0). \tag{17}$$

In words, we seek the largest amount of robustness such that our targeting still achieves $1-\alpha$ of the optimal procedure under the nominal scenario (no heterogeneity). Similar ideas have been used throughout decision-analysis. and there is growing empirical evidence that such models better capture how real decision-makers think (see, e.g., Brown and Sim (2009) and references therein).

In some sense we have simply replaced the problem of specifying λ with the problem of specifying α. However, from a practical point of view, we believe it is much more natural for a decision-maker to specify that she is willing to give up, say 10%, of revenues in the nominal scenario to protect herself against potential heterogeneity, than it is for her to specify a value for the coefficient of variation of the unknown heterogeneous causal effect in the candidate population. This is the perspective we take in our case study, and specify $\alpha = 10\%$.

Problem (17) can be solved by bisection search over λ. (See Theorem EC.2 in Appendix.)

4.8. Extensions of the Base Model

Appendix EC.2 considers several extensions to our base robust model and shows how one can naturally incorporate fairness considerations, domain-specific knowledge of the candidate CATE, evidence from multiple papers, and other generalizations.

5. Case Study: Comparison of Targeting Methods

Using data from our partner hospital, we seek to answer the following two questions: 1) when do robust methods outperform scoring rules, and 2) what drives this performance? We target a subset of Medicaid patients for case management at the end of 2014, with the goal of reducing the underpayments for ED charges from 1/1/2015 to 6/30/2015.

At the time of writing, there has not yet been a large-scale, landmark study quantifying the heterogeneous causal effects of case management. Hence, we do not have detailed CATE estimates, i.e., we do not have a "ground truth" against which to evaluate our methods. Our best empirical evidence to date is from Shumway et al. (2008) (Table 1). Using these data, we adopt the following approach to compare different methods:

1. We approximate a confidence interval for the SATE (Table 1). Had a researcher run a simple randomized control trial using data from Shumway et al. (2008), instead of a stratified analysis, she might have reported this estimate. This estimate is what would be typically available in a (non-stratified) published study.
2. We use this approximate confidence interval and summary statistics (Table 2) to compute our robust targeting solutions for two different variations of our uncertainty set (described below) corresponding to different (possibly misspecified) structures of CATEs.
3. We compare the performance of various methods in two different settings:

a. Section 5.4: We assume that the ground-truth candidate CATE is given by the estimates in Shumway et al. (2008). This setting is most relevant if we believe that the strata based on previous ED visits capture most of the heterogeneity in causal effects.

b. Section 5.5: We assume that the ground-truth candidate CATE depends only on demographic-related covariates: gender, race and age. This setting is most relevant if we believe that these demographics capture most of the heterogeneity in causal effects. A drawback of this setting is that we have no experimentally validated estimates of CATEs. Thus, we must focus on the worst-case performance.

Before proceeding to the details, we summarize our main findings:

• As predicted by Theorem 1, when treatment is benign or causal effects are nearly homogeneous, reward scoring performs well. However, this performance degrades rapidly if the treatment may be harmful. As the heterogeneity in CATE increases, scoring rules can be worse than not targeting.

• Our robust method performs almost as well as scoring rules when the heterogeneity (as measured by the adjusted CV) is small and much better than scoring rules as the adjusted CV increases, provided that the study CATE is approximately linear in the chosen description functions. If not, robust methods may perform poorly. These results agree qualitatively with Corollary 5.

• In this particular dataset, the benefits of our robust approach increase as the distribution of r_c has a shorter right tail or as the resource constraint K/C is tightened.

5.1. The Candidate Population

The data from our partner hospital include information on all adult Medicaid ED visits from 1/1/2014 to 12/31/2014. For each visit, we have the date of visit, total charges, associated primary ICD-10-CM diagnosis code, and patient identifier as well as the patient's demographic information, including age, gender, race and Medicaid eligibility. For patients who visited the ED in 2014, we also have the number of ED visits from 1/1/2015 to 6/30/2015. Table 3 provides the summary statistics for the patients before and after we apply Shumway et al. (2008)'s inclusion/exclusion criteria and limit our attention to Medicaid patients only. (See notes of Table 3 for detailed inclusion/exclusion criteria.) Despite applying the same inclusion-exclusion criteria, the resulting candidate population differs significantly from both the Medicaid ED patient population and the study population. (See Table 2 for a comparison of candidate and study populations.)

5.2. Implementation Details of Targeting Methods

For a fair comparison, all methods are assessed against the same values of r_c. Except in Section 5.6, these values are the average charges per ED visit for each patient in 2014. In Section 5.6, we assess the methods against other values of r_c with different tail behaviors. To maintain anonymity, we normalize performance metrics in dollar amounts by the full-information optimal value when it

Table 3 Inclusion Criteria and Summary Statistics for the Candidate Population

	All Medicaid ED Patients (24,943 **patients in** 2014)	**Candidate Population** ($C = 951$ **patients**)
Inclusion Criteria[§]		
Age, mean ± sd	43.3 ± 9.5	38.3 ± 12.5
No. of ED Visits in 2014		
1 - 4	23,558 (94%)	0 (0%)
5 - 11	1,286 (5%)	860 (90%)
≥ 12	99 (1%)	91 (10%)
Comorbidity [**]		
Alcohol Abuse	1,186 (5%)	777 (18%)
Drug Abuse	188 (0.8%)	39(5%)
Psychological Problem	1144 (5%)	216 (23%)
Medicaid Type[***]		
HUSKY A	8,767 (35%)	125 (13%)
HUSKY B	43 (0.2%)	0 (0%)
HUSKY C	1,697 (7%)	204 (21%)
HUSKY D	9,039 (36%)	695 (73%)
Other Characteristics		
Male	11,349 (45%)	442 (46%)
Race/Ethnicity		
African American	9,153 (37%)	475 (50%)
Hispanic	177 (1%)	7 (1%)
White	7,532 (30%)	283 (29%)
Other	8,081 (32%)	186 (20%)
Length of Stay (hours), mean ± sd	5.1 ± 7.4	6.2 ± 5.5
No. of ED Visits from 1/1/2015 to 6/30/2015 $y_c(0)$, mean ± sd	1.8 ± 1.7	7.2 ± 4.2
Avg. Charges Per ED Visit in 2014 r_c ($), mean ± sd	3,252 ± 4,491	3,324 ± 3,048
Charlson Comorbidity Score[‡], mean ± sd	0.08 ± 0.4	1.9 ± 0.9
Most Frequent Diagnosis[†] during ED Visits	back problems (5%)	alcohol-related disorders (10%)
	nonspecific chest pain (5%)	abdominal pain (6%)
	skin diseases (4%)	back problems (5%)
	upper respiratory infection (4%)	nonspecific chest pain (4%)
	alcohol-related disorders (4%)	connective tissue diseases (3%)
	sprains and strains (4%)	non-traumatic joint disorders (3%)

Notes. [§] Patients are included in the candidate population if they are at least 18 and below 65 years old; had at least 5 visits to the ED of our partner hospital in 2014; have a history of alcohol abuse, drug abuse, or psychological problems; have disability or blindness (HUSKY C); or have a low income and no dependent child (HUSKY D). [**] Calculated from the primary ICD-10-CM diagnosis code using Elixhauser Comorbidity Software (Elixhauser et al. 1998). [***] In Connecticut, Medicaid patients are eligible for one of the four parts (`http://www.ct.gov/hh/cwp/view.asp?a=3573&q=421548`). HUSKY A covers children, their parents and pregnant women; HUSKY B covers children whose parents earn too much money to qualify for Medicaid; HUSKY C covers low-income patients with disabilities or blindness; and HUSKY D covers the lowest-income patients with no dependent child. [‡] Calculated from the primary ICD-10-CM diagnosis code using Charlson Comorbidity Score (Charlson et al. 1987). [†] Calculated from the primary ICD-10-CM diagnosis code using Clinical Classification Software (Elixhauser et al. 2014).

is available (Section 5.4) or by the summation of K largest r_c when it is not (Section 5.5). Except for Section 5.6, we fix $K = 200$ ($K/C = 21\%$).

We compare the following four methods:

Reward Scoring (r-Scoring) We score by r_c.

Outcome Scoring ($ry(0)$-Scoring) We score by $r_c y_c(0)$, where $y_c(0)$ is the actual number of ED visits for patient c from 1/1/2015 to 6/30/2015.

Robust-2 We solve Problem (15) using a partition description function for strata. (Equivalently, we add the constraints $\beta_g = 0$ to all other description functions.) The corresponding summary statistics, i.e., the proportion of patients in each stratum, are given by Table 1.

Robust-Full-Linear We solve Problem (16) using partition description functions for strata, gender, and race and a linear description function for age. The corresponding summary statistics, i.e., the mean and standard deviation of each covariate, are given by Table 2.

We focus on reward and outcome scoring because, as mentioned, these methods are optimal when the causal effect is homogeneous and additive or multiplicative, respectively. Moreover, outcome scoring, in particular, closely mirrors current state-of-practice for targeting at our partner hospital.

We choose the adjusted CV parameter $\hat{\Gamma}_1/(\underline{I} - \hat{\Gamma}_2 - \hat{\kappa})$ via the satisficing heuristic in Section 4.7 with $\alpha = 10\%$, yielding 0.55 and 0.3 for Robust-2 and Robust-Full-Linear, respectively. In other words, we are willing to trade-off 10% of the cost saving in a nominal case for robustness. Finally, robust optimization problems frequently exhibit multiple optimal solutions. In our case study, we use the Pareto robust optimal solution corresponding to the realization $\Psi_{\mathsf{C}}(\cdot) = \underline{I} - \hat{\Gamma}_2 - \hat{\kappa}$, which can be computed as in Iancu and Trichakis (2013). Intuitively, this solution is non-dominated among robust optimal solutions when all patients respond to treatment identically.

5.3. Properties of the Solutions

Both Robust-2 and Robust-Full-Linear target all $K = 200$ patients when specifying their Adj. CV parameters as described above. This is essentially because targeting fewer than 200 patients would amount to more than a 10% loss in the nominal scenario. Indeed, as seen in Fig. EC.3 in Section EC.3.3, as long as one insists on less than a 50% loss in the nominal scenario, both robust methods fully utilize the budget.

Because of their different choices of description functions, the two robust methods match the distribution of covariates in the study population differently. Robust-2 attempts to match the proportion of patients in each stratum. Specifically, Table 4 shows that there are only 48% patients in stratum 1 in the study population, in contrast to 90% of the reward-weighted proportion of patients in stratum 1 in the candidate population. Thus, Robust-2 targets proportionally fewer stratum 1 patients, yielding an overall percentage of 78%. Notice that although this proportion is closer to the study population's 48%, it is not an exact match since the robust method also balances the competing objective of targeting higher-reward patients (recall Remark 4.5). In our candidate population, stratum 2 patients typically have lower r_c than stratum 1 patients (Figure EC.4 in the e-companion). Completely matching the proportion of stratum 1 patients would entail a significant loss in rewards. Also, although Robust-2 improves the matching of proportion of patients in each stratum, it exacerbates differences in other covariates, such as the proportion of Hispanic patients.

Similar observations can be made about Robust-Full-Linear. It more closely matches the means of the covariates in the study population than in the candidate population but does not always achieve exact matching. For example, our candidate population has very few Hispanic patients, so Robust-Full-Linear is unable to fully match the 22% of Hispanic patients in the study population. For all other covariates, it achieves a reasonably close match.

Table 4 Characteristics and Reward-Weighted Average Covariates for the Targeted Patients by Method

	Study Population	Candidate Population	Reward Scoring	Outcome Scoring	Robust-2	Robust-Full-Linear
Characteristics of Targeted Patients						
Avg. Charges Per ED Visit in 2014 r_c ($)		3,324	7,547	5,965	6,795	6,815
Avg. No. of ED Visits from 1/1/2015 to 6/30/2015 $y_c(0)$		2.3	2.2	5.3	3.0	2.6
Weighted Avg. Pre-Treatment Covariates*						
Demographics						
Male	75%	48%	51%	52%	53%	62%
African American	54%	46%	43%	44%	41%	53%
Hispanic	22%	0.5%	0.3%	0.1%	0.1%	0.3%
White	13%	34%	41%	42%	43%	44%
Age	43.3	40.7	44.9	44.9	45.3	44.3
Two Strata from Shumway et al. (2008)						
Stratum 1: 5 - 11 ED Visits in 2014	48%	90%	90%	83%	78%	84%

Note. * Except for the study population, we show the reward-weighted average of pre-treatment covariates. Thus, the reward-weighted summary statistics for the candidate population are different from those in Table 3.

5.4. When CATEs Depend Only on Previous ED Visits

In this section, the ground-truth candidate CATE is given by

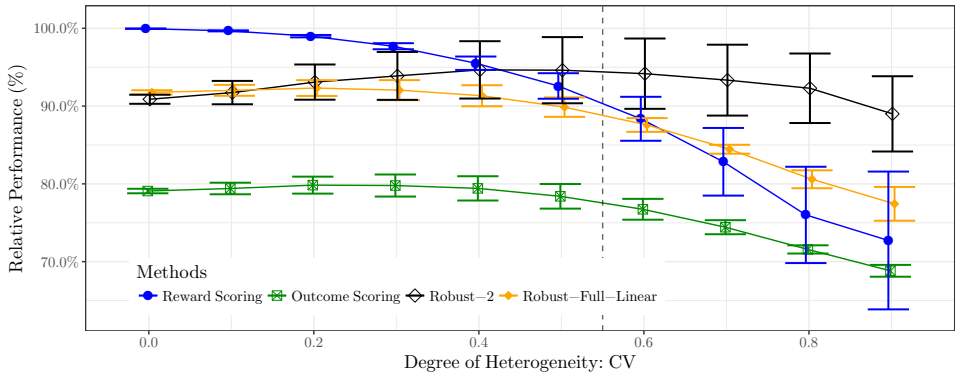
$$\mathbb{E}[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c] \equiv \psi_1 \cdot \mathbb{I}(\text{patient } c \text{ in stratum 1}) + \psi_2 \cdot \mathbb{I}(\text{patient } c \text{ in stratum 2}) \tag{18}$$

for all $c \in \{1, \ldots, C\}$, where ψ_1 and ψ_2 denote the true CATEs for stratum 1 (5 to 11 ED visits in the previous year) and stratum 2 (greater than or equal to 12 ED visits in the previous year).

For the particular point estimates of $\psi_1 = 2.1$, $\psi_2 = 3.3$ (ED visits reduced) provided by Shumway et al. (2008), reward scoring achieves 99% of the full-information optimum benchmark, while Robust-2 achieves 95% and Robust-Full-Linear achieves 93%.

However, the point estimates for ψ_1, ψ_2 are not exact, so our assessment might be optimistic. We next study the performance under small perturbations of these estimates. Specifically, we vary ψ_1, ψ_2 uniformly within the confidence intervals provided by Shumway et al. (2008): $[0.1, 4.1]$ and $[0.2, 6.4]$, respectively. Each pair of values yields a different relative performance for each method and a different level of heterogeneity in the candidate CATE, as measured by its coefficient of variation. This coefficient of variation is given by $\text{CV} = \sqrt{\sum_{g=1}^{2} \mu_g (\psi_g - \boldsymbol{\mu}^T \boldsymbol{\psi})^2} / \boldsymbol{\mu}^T \boldsymbol{\psi}$, where $\mu_1 =$

Figure 2 Relative Performance as Both ψ_1 and ψ_2 Vary Uniformly within Their Confidence Intervals



Note. Performance is relative to the full-information benchmark, and ψ_1, ψ_2 are the number of ED visits reduced in each stratum. We plot the mean relative performance (point) across choices of ψ_1, ψ_2 with the given CV, plus/minus one standard deviation (error bar). To the left of the dashed line CV$=0.55$, the ground-truth CATEs belong to the uncertainty set of the Robust-2 method.

0.48 and $\mu_2 = 0.52$ from Table 2. We summarize the relative performance of each method by plotting the mean and standard deviation for a given level of CV across all perturbations in Figure 2.

Perhaps unsurprisingly, all methods perform reasonably well, obtaining at least 65% of the full-information optimal value. Indeed, the confidence intervals of CATEs are both strictly positive, i.e., treatment is benign, and Theorem 1 ensures that reward scoring cannot be too suboptimal. Nonetheless, as the heterogeneity increases, we do observe qualitative differences in behavior.

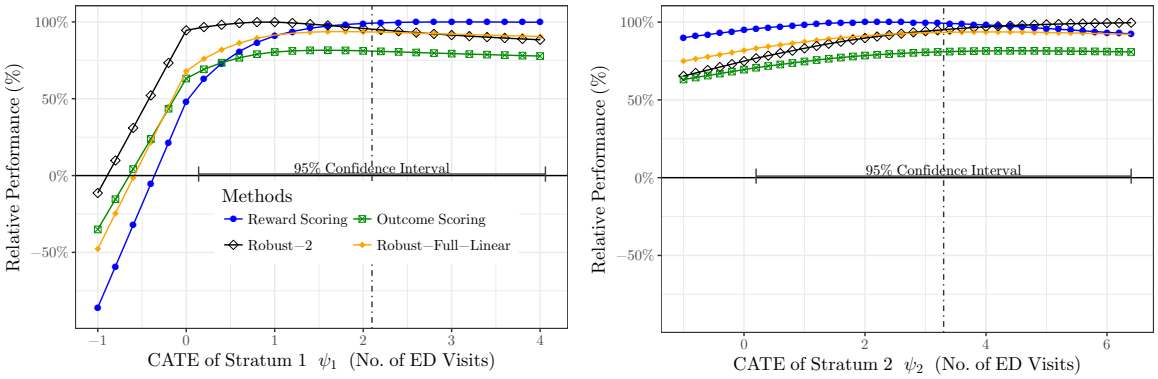
When CV$=0$, reward scoring is optimal (as expected by Corollary 1). As CV increases, reward scoring's performance degrades rapidly. At worst, it obtains about 75% relative performance. Robust-2 performs slightly worse than reward scoring when the degree of heterogeneity is small (obtaining about 90% relative performance) and much better than reward scoring as CV increases up to 0.4 (obtaining almost 95% relative performance). A similar observation can be made for Robust-Full-Linear: Robust-Full-Linear performs almost as well as reward scoring when the degree of heterogeneity is small and is better than reward scoring when it is sufficiently large.

Notice also that when CV < 0.55, the ground-truth CATEs in our simulation belong to the uncertainty set defined by Robust-2. However, because of its "worst-case perspective", Robust-2 does not outperform Reward Scoring unless the CV is sufficiently large.

One way to interpret the value CV $= .4$ (where the methods intersect) is that since SATE is approximately 2.7, CV $= .4$ implies the standard deviation of the candidate CATE is at most $.4 \cdot 2.7 = 1.08$. In other words given two random patients, the expected absolute difference between their causal effect is at most $\sqrt{2} \cdot 1.08 \approx 1.5$ visits.[5] Thus, if we believe that benefit of case management varies by more than 1.5 visits between patients, Robust-2 outperforms reward scoring.

[5] For X, Y i.i.d. random variables, $\mathbb{E}[|X-Y|] \leq \sqrt{\mathbb{E}[(X-Y)^2]} = \sqrt{2\mathrm{Var}(X)}$ by Jensen's inequality.

Figure 3 Relative Performance as Either ψ_1 or ψ_2 Varies



Note. Negative values of ψ_1, ψ_2 indicate increased ED visits. In the left panel, we fix $\psi_2 = 3.3$ and vary ψ_1, while in the right panel, we fix $\psi_1 = 2.1$, and vary ψ_2. The dashed vertical lines indicate the point estimates of the varying parameter.

To further investigate the effects of heterogeneity, we vary only one of ψ_1 or ψ_2 in Figure 3 and keep the other fixed at its point estimate from Shumway et al. (2008). The performance of reward scoring degrades rapidly as ψ_1 decreases, and it can perform very badly when ψ_1 becomes negative, i.e., when case management may increase ED visits for patients in stratum 1. To the contrary, both robust methods outperform reward scoring significantly in these instances. As mentioned in Section 5.3, both robust methods effectively target fewer stratum 1 patients, which makes their performance less sensitive to the changes in ψ_1. Of course, since these methods target more stratum 2 patients, they are more sensitive to the value of ψ_2 (see right panel of Figure 3), but the magnitude of the changes are substantially smaller. Overall, then, we would argue that the robust methods are indeed more "robust" to uncertainties in ψ_1 and ψ_2.

5.5. When CATEs Depend Only on Patient Demographics

In this section, the ground-truth study CATE is given by

$$\mathbb{E}[\delta^{\tilde{s}} \mid \mathbf{x}^{\tilde{s}} = \mathbf{x}^s] \equiv \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}^s) + \epsilon(\mathbf{x}^s), \quad \forall s \in \{1, \ldots, S\}, \tag{19}$$

for some $\beta_0, \boldsymbol{\beta}$ such that $\beta_g \neq 0$ only if g corresponds to a demographic-related covariate, i.e,. the patient's gender, race or age. Let $\mathcal{G}$ be the set of indices for these demographic-related description functions. Since we do not have experimentally validated estimates for a CATE with this structure, we will compare to worst-case performance over the uncertainty set

$$\mathcal{U}^*_{\boldsymbol{\Gamma},\kappa} = \left\{ \Psi_{\mathsf{C}} : \mathcal{X} \mapsto \mathbb{R} \;\middle|\; \exists \epsilon : \mathcal{X} \mapsto \mathbb{R},\ \beta_0 \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^G,\ \text{s.t.}\ \ \Psi^{\mathsf{S}}(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}) + \epsilon(\mathbf{x}),\ \beta_g = 0\ \forall g \notin \mathcal{G}, \right. \tag{20}$$

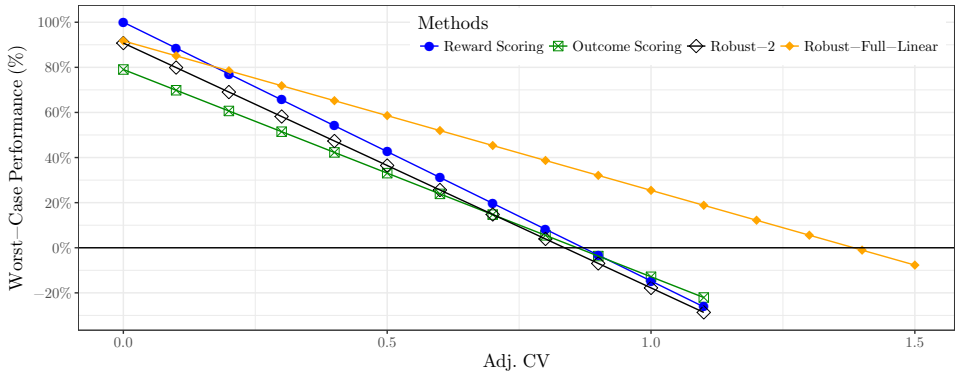
$$\left. \underline{I} \leq \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \leq \overline{I},\ \ \left\| \left(\Psi^{\mathsf{S}}(\mathbf{x}_c) - \Psi_{\mathsf{C}}(\mathbf{x}_c) \right)_{c=1}^{C} \right\|_{\infty} \leq \kappa,\ \ \|\boldsymbol{\beta}\|_{\mathbf{V}} \leq \Gamma_1,\ \ \left\| (\epsilon(\mathbf{x}^*_c))_{c=1}^{C} \right\|_{\infty} \leq \Gamma_2 \right\},$$

for varying κ, Γ_1, and Γ_2. Here, $\mathbf{V}$ is diagonal with the variance of demographic-related covariates as entries (Table 2).

Using Theorem 3, we can compute the worst-case objective in closed form. To maintain anonymity, we normalize the worst-case objective by $(\underline{I} - \Gamma_2 - \kappa)\sum_{c=1}^{K} r_c$, where $r_1 \geq \cdots \geq r_C$. The final worst-case performance metric only depends on the ratio Adj. CV $\equiv \Gamma_1/(\underline{I} - \Gamma_2 - \kappa)$. Inspired by this fact, we present our results relative to this "true" Adj. CV, and use the notation Adj. $\widehat{\text{CV}} \equiv \hat{\Gamma}_1/(\underline{I} - \hat{\Gamma}_2 - \hat{\kappa})$ for the parameter of the uncertainty set used to compute Robust-2, or Robust Full-Linear, depending on the chosen setting.

Note that both robust methods are "misspecified" under (19) and (20). Specifically, Robust-Full-Linear assumes that CATEs may depend on the strata membership (when they, in fact, do not) and assumes a particular value of Adj. $\widehat{\text{CV}}$, which, in general, is different from the true Adj. CV above. We also stress that the treatment in this experiment can be potentially harmful.

Figure 4 Worst-Case Performance When CATEs Depend Only on Demographics



Note. Worst-case performance is the worst-case objective value under uncertainty set (20) normalized by $(\underline{I} - \Gamma_2 - \kappa)\sum_{c=1}^{K} r_c$.

We plot the anonymized worst-case performance against Adj. CV for different methods in Figure 4. Consider reward scoring and Robust-Full-Linear. When Adj. CV $= 0$, reward scoring is optimal. Notice that Robust-Full-Linear performs not optimally but at least 90% compared to scoring rule due to our proposed heuristics of choosing Adj. $\widehat{\text{CV}}$ in Section 4.7. However, the performance of reward scoring degrades rapidly as Adj. CV increases, while Robust-Full-Linear performs significantly better than scoring rules as Adj. CV increases above 0.2.

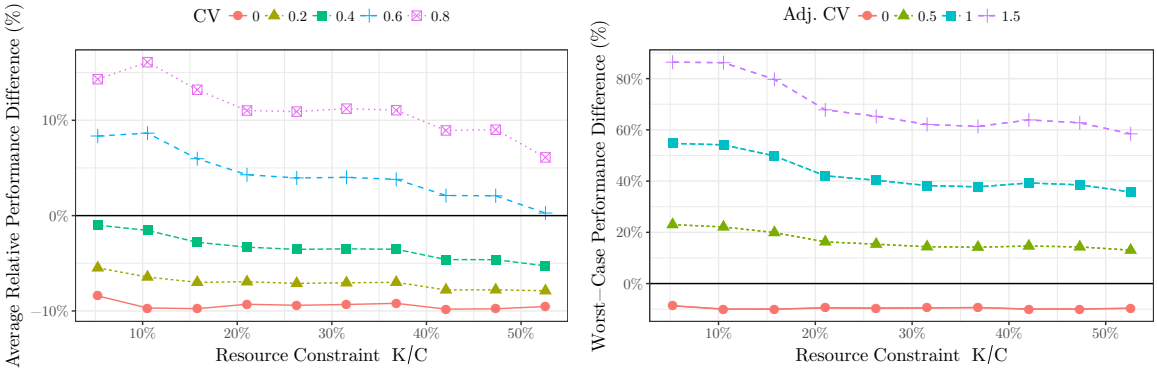
Robust-2 performs poorly in this experiment. The worst-case performance of Robust-2 can be negative and worse than not targeting. We partially explain this behavior using Corollary 5 in Section EC.3.4 of the appendix. Intuitively, the variation in the true CATE is not well-captured by the description functions in Robust-2. Consequently, Γ_2 must be very large before the true CATE is contained in its uncertainty set, making the uncertainty set very large and the worst-case performance very poor. In other words, Robust-2 is highly misspecified.

5.6. Sensitivity Analysis

Reward scoring performs well compared to robust methods when the treatment is benign. We argued that this is due to the long-tail behavior of the reward distribution. In this section, we seek to verify this by exploring how the performance of our robust methods and reward scoring changes as we vary the tail behavior of r_c distribution and the resource constraint K/C.

We only include results for Robust-Full-Linear and reward scoring, but we observe similar behavior for Robust-2 and outcome scoring (see Appendix EC.3.5). Throughout, we focus on the performance difference between Robust-Full-Linear and reward scoring under the ground-truth models of Sections 5.4 and 5.5. For each setting, we compute the performance metrics for the robust method and reward scoring separately and report their difference. For example, under the ground truth of Section 5.4, we compute the average relative performance to the full-information optimum for each method and determine the difference. We refer to this quantity as the *average relative performance difference*. Under the ground truth of Section 5.5, we compute the worst-case performance of Robust-Full-Linear and reward scoring (normalized by $(\underline{I} - \Gamma_2 - \kappa)\sum_{c=1}^{K} r_c$) separately and determine the difference. We term this quantity the *worst-case performance difference*. A positive performance difference implies that Robust-Full-Linear outperforms reward scoring.

Figure 5 Difference between Robust-Full-Linear and Reward Scoring Varying the Resource Constraint



Note. The left panel corresponds to Figure 2 in Section 5.4. The right panel corresponds to Figure 4 Section 5.5. In both panels, we plot the performance difference (defined in the main text) against K/C.

To investigate the sensitivity of resource constraint, we vary the value of K and reproduce the experiments of Figs. 2 and 4 in terms of the performance difference between Robust-Full-Linear and reward scoring in Figure 5. For each value of K, we use our satisficing approach of Section 4.7 with $\alpha = 10\%$ to specify the parameters of the uncertainty set. For all values of K studied, i.e., $K = 10, 20, \ldots, 50, 100, 150, \ldots, 800$, both Robust Full-Linear and Robust-2 target all K patients. For brevity, we only present results for Robust Full-Linear.

We see that the performance difference increases as the resource constraint is tightened, i.e., as K/C decreases, in both settings. In addition, for any fixed level of K/C, Robust-Full-Linear tends

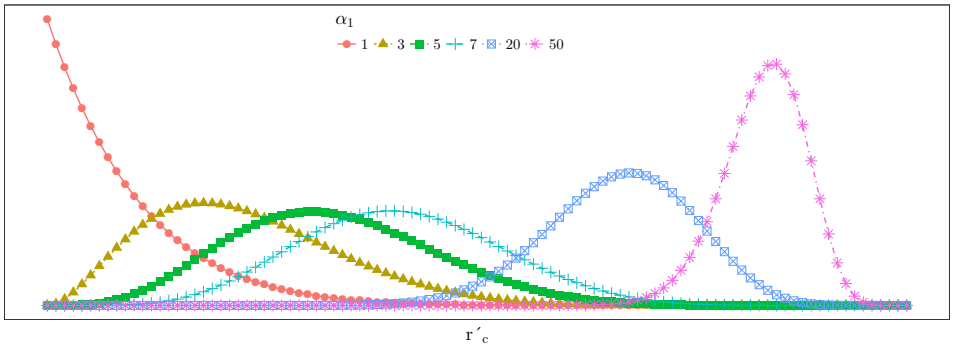
to increasingly outperform reward scoring as CV or Adj. CV increases, which is consistent with our previous observations.

To investigate the sensitivity to the reward distribution, we reproduce the experiments in Figure 2 and Figure 4 in terms of performance differences for a variety of different reward distributions. Specifically, define the new rewards

$$r'_c \equiv F^{-1}_{\alpha_1,10}(\hat{F}(r_c)), \quad \forall c = 1, \ldots, C, \tag{21}$$

where $\hat{F}(\cdot)$ is the empirical cumulative distribution function (CDF) of original rewards r_c (average charges per ED visit for patient c in 2014) and $F^{-1}_{\alpha_1,10}(\cdot)$ is the inverse CDF of beta distribution with parameter α_1 and $\alpha_2 = 10$. By varying α_1, we can alter the tail behavior of the reward distribution (see Figure 6). Although this transformation may change the average reward in the candidate population, the scale of rewards does not affect our performance metrics.

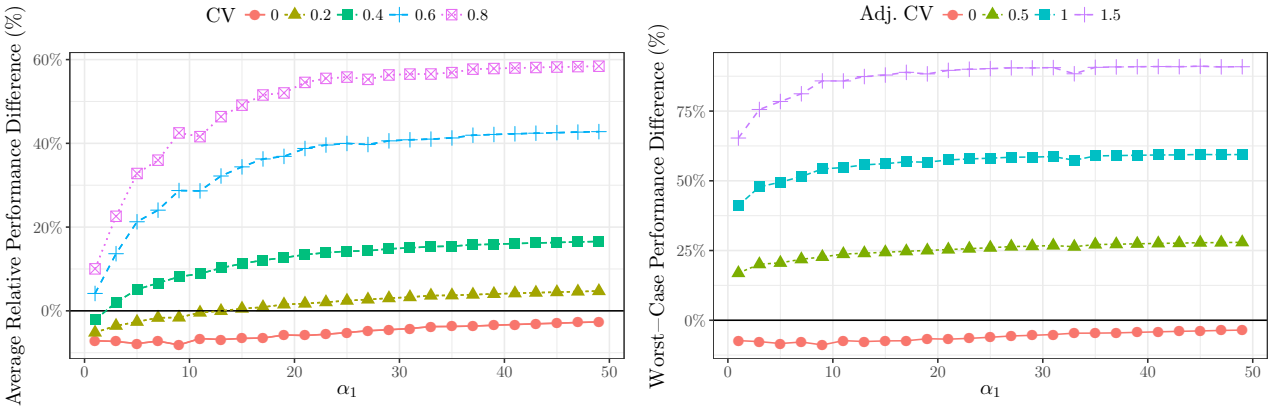
Figure 6 Distribution of Rewards r'_c defined in Eq. (21)



Note. The parameter $\alpha_2 = 10$ and α_1 varies.

For varying α_1, we re-compute rewards by Eq. (21), re-compute each method and assess the performance difference with the new rewards under our two possible ground truths (Figure 7).

Figure 7 Difference between Robust-Full-Linear and Reward Scoring Varying Reward Distribution



Note. The left panel corresponds to Figure 2 in Section 5.4. The right panel corresponds to Figure 4 Section 5.5.

Similar to previous experiments, for a fixed value of α_1, we see an increasing performance gap as CV or Adj. CV increases. More interestingly, the performance difference increases as α_1 increases, although with severely decreasing marginal returns; for large values of α_1, there is almost no benefit. This effect is significantly less pronounced in the right panel when considering worst-case behavior. Overall, it seems that for this particular dataset, the benefits of the Robust-Full-Linear method over reward scoring increase as the reward distribution has a shorter right tail. We write "for this dataset" since it is possible to construct datasets where this finding is not true.

This result agrees well with our previous intuition that targeting patients with very high rewards is "risky" since the performance will depend strongly on the unknown causal effect of these high-reward patients. Intuitively, if case-management causes these high-reward patients to visit the ED more frequently, our effectiveness decreases substantially. In the left panel, reward scoring performs well (i.e., the performance difference is negative) for the initial reward distribution at least partially because the highest-reward patients mostly belong to stratum 1 and because these patients have the highest causal effects. As the reward distribution shifts, the difference in rewards between stratum 1 and stratum 2 patients shrinks, so this benefit erodes, and the performance difference becomes positive. By contrast, in the right-hand panel, since we consider worst-case behavior, reward scoring does not enjoy such a benefit under the initial reward distribution, and we see a much smaller gain as we shift the distribution.

6. Conclusion

We proposed a robust optimization model to maximize intervention effectiveness utilizing evidence available from published studies. Our approach is intuitive, flexible, and computationally tractable and outperforms current practice when the underlying heterogeneity in causal effects are large.

Acknowledgement: We would like to thank the Department Editor, Associate Editor and the two anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

References

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research* 55(1):80–98.

Athey S, Imbens GW (2015) Machine learning methods for estimating heterogeneous causal effects. *Stat* 1050(5).

Athey S, Wager S (2017) Efficient policy learning. *arXiv preprint arXiv:1702.02896* .

Bastani H, Bayati M (2016) Online decision-making with high-dimensional covariates. *SSRN: https://ssrn.com/abstract=2661896* .

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Ben-Tal A, Den Hertog D, Vial JP (2015) Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming* 149(1-2):265–299.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton, New Jersey: Princeton University Press).

Bertsimas D, Copenhaver MS (2017) Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research* .

Bertsimas D, Gupta V, Kallus N (2017) Robust sample average approximation. *Mathematical Programming* 1–66.

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Mathematical Programming* 167(2):235–292.

Bertsimas D, O'Hair A, Relyea S, Silberholz J (2016a) An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science* 62(5):1511–1531.

Bertsimas D, Silberholz J, Trikalinos T (2016b) Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening. *Health Care Management Science* 1–14.

Billings J, Parikh N, Mijanovich T (2000) Emergency department use in New York City: A substitute for primary care? *Issue Brief (Commonwealth Fund)* 433:1–5.

Billings J, Raven MC (2013) Dispelling an urban legend: Frequent emergency department users have substantial burden of disease. *Health Affairs* 32(12):2099–2108.

Bortfeld T, Chan TCY, Trofimov A, Tsitsiklis JN (2008) Robust management of motion uncertainty in intensity-modulated radiation therapy. *Operations Research* 56(6):1461–1473.

Brown DB, Sim M (2009) Satisficing measures for analysis of risky positions. *Management Science* 55(1):71–84.

Chan TCY, Demirtas D, Kwon RH (2016) Optimizing the deployment of public access defibrillators. *Management Science* 62(12):3617–3635.

Chan TCY, Shen ZJM, Siddiq A (2017) Robust defibrillator deployment under cardiac arrest location uncertainty via row-and-column generation. *Operations Research* .

Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases* 40(5):373–383.

Cole SR, Stuart EA (2010) Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology* 172(1):107–115.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612.

Deming D, Dynarski S (2009) Into college, out of poverty? Policies to increase the postsecondary attainment of the poor. Technical report, National Bureau of Economic Research.

Den Hertog D (2018) Is DRO the Only Approach for Optimization Problems with Convex Uncertainty? URL `https://www.birs.ca/events/2018/5-day-workshops/18w5102/videos/watch/201803080905-denHertog.html`.

Deo S, Rajaram K, Rath S, Karmarkar US, Goetz MB (2015) Planning for HIV screening, testing, and care at the Veterans Health Administration. *Operations Research* 63(2):287–304.

Eichler HG, Abadie E, Breckenridge A, Leufkens H, Rasi G (2012) Open clinical trial data for all? A view from regulators. *PLoS Medicine* 9(4):e1001202.

Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. *Medical Care* 36(1):8–27.

Elixhauser A, Steiner C, Palmer L (2014) Clinical Classifications Software (CCS). Agency for Healthcare Research and Quality, URL `https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp`.

Esfahani PM, Kuhn D (2015) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 1–52.

Gao R, Chen X, Kleywegt AJ (2017) Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050* .

Ghaoui LE, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications* 18(4):1035–1064.

Gibbs AL, Su FE (2002) On choosing and bounding probability metrics. *International statistical review* 70(3):419–435.

Goh J, Bayati M, Zenios SA, Singh S, Moore D (2018) Data uncertainty in markov chains: Application to cost-effectiveness analyses of medical innovations. *Operations Research* .

Gutierrez P, Gérardy JY (2017) Causal inference and uplift modelling: A review of the literature. *International Conference on Predictive Applications and APIs*, 1–13.

Hartman E, Grieve R, Ramsahai R, Sekhon JS (2015) From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects. *Journal of Royal Statistical Society: Series A (Statitsics in Society)* 10:1111.

Higgins J, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21(11):1539–1558.

Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–960, ISSN 01621459, URL `http://www.jstor.org/stable/2289064`.

Iancu DA, Trichakis N (2013) Pareto efficiency in robust optimization. *Management Science* 60(1):130–147.

Imai K, Ratkovic M, et al. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1):4–29.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Jackson C, DuBard A (2015) It's all about impactability! Optimizing targeting for care management of complex patients. *Community Care of North Carolina.* Data Brief 4, URL `https://www.communitycarenc.org/media/files/data-brief-no-4-optimizing-targeting-cm.pdf`.

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

Kallus N (2017) Recursive partitioning for personalization using observational data. Precup D, Teh YW, eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1789–1798 (International Convention Centre, Sydney, Australia: PMLR).

Kuhn D, Wiesemann W, Georghiou A (2011) Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming* 130(1):177–209.

Lam H (2016) Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research* 41(4):1248–1275.

Lee KH, Davenport L (2006) Can case management interventions reduce the number of emergency department visits by frequent users? *The Health Care Manager* 25(2):155–159.

Negoescu DM, Bimpikis K, Brandeau ML, Iancu DA (2017) Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science* .

Pardo L (2005) *Statistical inference based on divergence measures* (Chapman and Hall/CRC).

Phillips GA, Brophy DS, Weiland TJ, Chenhall AJ, Dent AW (2006) The effect of multidisciplinary case management on selected outcomes for frequent attenders at an emergency department. *Medical Journal of Australia* 184(12):602.

Shah R, Chen C, O'Rourke S, Lee M, Mohanty SA, Abraham J (2011) Evaluation of care management for the uninsured. *Medical Care* 49(2):166–171.

Shumway M, Boccellari A, O'Brien K, Okin RL (2008) Cost-effectiveness of clinical case management for ED frequent users: Results of a randomized trial. *The American Journal of Emergency Medicine* 26(2):155–164.

Simon HA (1955) A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1):99–118.

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386.

Vizcaíno VM, Aguilar FS, Gutiérrez RF, Martínez MS, López MS, Martínez SS, García EL, Artalejo FR (2008) Assessment of an after-school physical activity program to prevent obesity among 9 to 10-year-old children: A cluster randomized trial. *International Journal of Obesity* 32(1):12.

Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *Journal of Machine Learning Research* 10(Jul):1485–1510.

Zhao Y, Fang X, Simchi-Levi D (2017) A practically competitive and provably consistent algorithm for uplift modeling. *Data Mining (ICDM), 2017 IEEE International Conference on*, 1171–1176 (IEEE).

Zhen J, de Ruiter FJ, den Hertog D (2017a) Robust optimization for models with uncertain SOC and SDP constraints. *Optimization Online PrePrint* URL `http://www.optimization-online.org/DB_HTML/2017/12/6371.html`.

Zhen J, den Hertog D, Sim M (2017b) Adjustable robust optimization via fourier-motzkin elimination. *Operations Research* .

Zuckerman S, Williams AF, Stockley KE (2009) Trends in medicaid physician fees, 2003–2008. *Health Affairs* 28(3):w510–w519.

Proofs and Additional Graphs

EC.1. Proofs

Proof of Theorem 1. We require the following identity:

$$\frac{v+w}{v+z} \geq \frac{\underline{v}+w}{\underline{v}+z}, \quad \text{whenever } v \geq \underline{v},\ z \geq w. \tag{EC.1}$$

To prove the identity, differentiate the left-hand side by v, yielding

$$\frac{z-w}{(v+z)^2} \geq 0,$$

since $z \geq w$ by assumption. Thus, decreasing v to $\underline{v}$ never increases the ratio in (EC.1), which proves the identity.

Recall that B^* denotes the full-information optimal solution to Problem (1). We assume that $|B^*| = K$ without loss of generality. Since $\delta_c/\hat{\delta}_c \geq \underline{\delta} > 0$, we have $r_c\delta_c \geq 0$ for all $c \in \{1,\ldots,K\}$. If $|B^*| < K$, we can always target additional patients in $\{1,\ldots,K\}$ without decreasing the objective.

For notational convenience, let $a_c \equiv r_c\hat{\delta}_c$ and $b_c \equiv \frac{\delta_c}{\hat{\delta}_c}$ for all $c \in \{1,\ldots,C\}$. Then, $r\hat{\delta}$-scoring is equivalent to a-scoring, and the objective of Problem (1) can be written as $\sum_{c=1}^{C} r_c\delta_c = \sum_{c=1}^{C} a_c b_c$.

The relative performance of the a-scoring rule is

$$\frac{\sum_{c:1\leq c\leq K} a_c b_c}{\sum_{c:c\in B^*} a_c b_c} = \frac{\sum_{c:1\leq c\leq K, c\in B^*} a_c b_c + \sum_{c:1\leq c\leq K, c\notin B^*} a_c b_c}{\sum_{c:1\leq c\leq K, c\in B^*} a_c b_c + \sum_{c:c\geq K+1, c\in B^*} a_c b_c}. \tag{EC.2}$$

Since B^* is the full-information optimal solution,

$$\sum_{c:c\in B^*} a_c b_c \geq \sum_{c:1\leq c\leq K} a_c b_c \quad \Leftrightarrow \quad \sum_{c:c\geq K+1, c\in B^*} a_c b_c \geq \sum_{c:1\leq c\leq K, c\notin B^*} a_c b_c.$$

By assumption, we also have $b_c \geq \underline{\delta}$ for all $c \in \{1,\ldots,K\}$. Applying the identity in (EC.1) yields

$$\frac{\sum_{c:1\leq c\leq K} a_c b_c}{\sum_{c:c\in B^*} a_c b_c} \geq \frac{\underline{\delta}\sum_{c:1\leq c\leq K, c\in B^*} a_c + \sum_{c:1\leq c\leq K, c\notin B^*} a_c b_c}{\underline{\delta}\sum_{c:1\leq c\leq K, c\in B^*} a_c + \sum_{c:c\geq K+1, c\in B^*} a_c b_c}. \tag{EC.3}$$

Then,

$$\text{Eq.(EC.3)} \ \geq \ \frac{\underline{\delta}\sum_{c:1\leq c\leq K} a_c}{\underline{\delta}\sum_{c:1\leq c\leq K, c\in B^*} a_c + \overline{\delta}\sum_{c:c\geq K+1, c\in B^*} a_c}, \tag{EC.4}$$

since $b_c \geq \underline{\delta}$ for all $c \in \{1,\ldots,K\}$. Define $k \equiv |B^* \cap \{1,\ldots,K\}|$ to be the number of patients targeted by both methods. Then,

$$\begin{aligned}\text{Eq. (EC.4)} \ &\geq \ \frac{\underline{\delta}\sum_{c=1}^{K} a_c}{\underline{\delta}\sum_{c=1}^{k} a_c + \overline{\delta}\sum_{c=K+1}^{2K-k} a_c} \quad \text{(since } a_c \text{ is in decreasing order and } \underline{\delta} > 0\text{)} \\ &\geq \ \frac{\underline{\delta}\sum_{c=1}^{K} a_c}{\max_{0\leq k\leq K}\{\underline{\delta}\sum_{c=1}^{k} a_c + \overline{\delta}\sum_{c=K+1}^{2K-k} a_c\}}. \end{aligned} \tag{EC.5}$$

Consider the maximization in (EC.5) and rewrite the summations:

$$\underline{\delta}\sum_{c=1}^{k} a_c + \overline{\delta}\sum_{c=K+1}^{2K-k} a_c \;=\; \overline{\delta}\sum_{c=K+1}^{2K} a_c + \sum_{c=1}^{k}[\underline{\delta}a_c - \overline{\delta}a_{2K-c+1}].$$

The first term does not depend on k. In the second term, note that for $c < c'$, we have $\underline{\delta}a_c - \overline{\delta}a_{2K-c+1} \geq \underline{\delta}a_{c'} - \overline{\delta}a_{2K-c'+1}$ since a_c is in descending order and $\underline{\delta}, \overline{\delta} > 0$. Therefore, the quantity $\underline{\delta}a_c - \overline{\delta}a_{2K-c+1}$ is also decreasing in c. It follows that the optimal k^* is the largest c such that $1 \leq c \leq K$ and $\underline{\delta}a_c - \overline{\delta}a_{2K-c+1} \geq 0$ and is 0 if this quantity is always non-positive. This proves the inequality (2).

We next give an example in which the bound is achieved. Take

$$\delta_c = \begin{cases} \underline{\delta}\hat{\delta}_c, & \text{if } 1 \leq c \leq K \\ \overline{\delta}\hat{\delta}_c, & \text{otherwise,} \end{cases} \quad \Leftrightarrow \quad b_c = \begin{cases} \underline{\delta}, & \text{if } 1 \leq c \leq K \\ \overline{\delta}, & \text{otherwise.} \end{cases}$$

For these values, we will confirm that the true relative performance of a-scoring is given by (2). We first show that $B^* = \{1, \ldots, k^*\} \cup \{K+1, \ldots, 2K-k^*\}$, where k^* is defined in the theorem. To see this, note that since a_c is non-increasing and b_c is constant on the scale $1 \leq c \leq K$, B^* must be of the form $\{1, \ldots, k\} \cup \{K+1, 2K-k\}$ for some $0 \leq k \leq K$, and the full-information objective value is $\max_{0 \leq k \leq K}\{\underline{\delta}\sum_{c=1}^{k} a_c + \overline{\delta}\sum_{c=K+1}^{2K-k} a_c\}$. As proven previously, k^* optimizes this objective so that B^* has the required form. Substituting in the definition of a_c, b_c and B^* into Eq. (EC.2) completes the proof. □

Proof of Corollary 1. For the first part, take the derivative of $\omega(\underline{\delta}/\overline{\delta})$ with respect to $\underline{\delta}/\overline{\delta}$ to obtain

$$\omega'(\underline{\delta}/\overline{\delta}) = \frac{\sum_{c=1}^{K} r_c\hat{\delta}_c \sum_{c=1}^{k^*} r_c\hat{\delta}_c}{((\underline{\delta}/\overline{\delta})\sum_{c=1}^{k^*} r_c\hat{\delta}_c + \sum_{c=K+1}^{2K-k^*} r_c\hat{\delta}_c)^2} \geq 0,$$

where the inequality follows because $r_c\hat{\delta}_c \geq 0$ for all $c \in \{1, \ldots, C\}$ by assumption.

For the second part, we apply the definition of k^* in Theorem 1. When $\underline{\delta}/\overline{\delta} \geq r_{K+1}\hat{\delta}_{K+1}/r_K\hat{\delta}_K$, $k^* = K$ and $\omega(\underline{\delta}/\overline{\delta}) = 1$.

Finally, when $(\underline{\delta}/\overline{\delta})$ is sufficiently small, $(\underline{\delta}/\overline{\delta}) \leq r_{2K}\hat{\delta}_{2K}/r_1\hat{\delta}_1$, and we have $k^* = 0$. Then, the denominator of ω does not depend on $\underline{\delta}/\overline{\delta}$, and the numerator goes to 0 as $\underline{\delta}/\overline{\delta} \to 0$, so $\omega(\underline{\delta}/\overline{\delta}) \to 0$. Thus, the proof is complete. □

Proof of Theorem 2. First note that $\mathbf{0}$ is feasible and has worst-case performance 0 in (8). Thus, it suffices to show that for any $\mathbf{z} \notin \mathcal{Z}^*$, the worst-case objective over (9) is at most 0. For such a $\mathbf{z}$, there must exist $\mathbf{x}_0 \in \mathcal{X}$ such that $q_{\mathbf{z}}(\mathbf{x}_0) \neq \mathbb{P}(\mathbf{x}^{\tilde{s}} = \mathbf{x}_0)$. Consider the function $\Psi_\lambda(\mathbf{x}) = \underline{I} + \lambda(\mathbb{I}(\mathbf{x} = \mathbf{x}_0) - \mathbb{P}(\mathbf{x}^{\tilde{s}} = \mathbf{x}_0))$. Since $\hat{\kappa} \geq 0$, $\Psi_\lambda \in \mathcal{U}_{\hat{\kappa}}$. Furthermore,

$$\sum_{c=1}^{C} z_c r_c \Psi_\lambda(\mathbf{x}_c) = \sum_{c=1}^{C} z_c r_c \left(\sum_{\mathbf{x} \in \mathcal{X}} q_{\mathbf{z}}(\mathbf{x})\Psi_\lambda(\mathbf{x}_c)\right)$$

$$= \sum_{c=1}^{C} z_c r_c \left(\underline{I} + \sum_{\mathbf{x} \in \mathcal{X}} q_{\mathbf{z}}(\mathbf{x}) \lambda (\mathbb{I}(\mathbf{x}_c = \mathbf{x}_0) - \mathbb{P}(\mathbf{x}^{\tilde{s}} = \mathbf{x}_0)) \right)$$
$$= \sum_{c=1}^{C} z_c r_c \left(\underline{I} + \lambda (q_{\mathbf{z}}(\mathbf{x}_0) - \mathbb{P}(\mathbf{x}^{\tilde{s}} = \mathbf{x}_0)) \right).$$

Now, if $\sum_{c=1}^{C} z_c r_c = 0$, this quantity is 0. Otherwise, if $\sum_{c=1}^{C} z_c r_c > 0$, then, by taking $\lambda \to \pm\infty$, we have that the worst-case performance over (9) of $\mathbf{z}$ is $-\infty$. In either case, the worst-case performance is at most 0. This proves the theorem. □

Proof of Theorem 3. We apply standard techniques (see, e.g., Ben-Tal et al. 2009 for a review). Given any feasible $\mathbf{z}$, the inner minimization of (8) under uncertainty set (11) can be rewritten as

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, (\epsilon(\mathbf{x}_c))_{c=1}^{C}, (v(\mathbf{x}_c))_{c=1}^{C}} \quad & \sum_{c=1}^{C} z_c r_c \left(\beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}_c) + \epsilon(\mathbf{x}_c) + v(\mathbf{x}_c) \right) \\ \text{s.t.} \quad & \underline{I} \le \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \le \overline{I}, \ \|\boldsymbol{\beta}\| \le \hat{\Gamma}_1 \\ & \|(\epsilon(\mathbf{x}_c))_{c=1}^{C}\|_{\mathsf{res}} \le \hat{\Gamma}_2, \ \|(v(\mathbf{x}_c))_{c=1}^{C}\|_{\mathsf{link}} \le \hat{\kappa}, \end{aligned}$$

where $v(\mathbf{x}_c)$ represents the difference $\Psi_{\mathsf{C}}(\mathbf{x}_c) - \Psi^{\mathsf{S}}(\mathbf{x}_c)$. This optimization problem decomposes into the sum of three separate minimizations:

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}} \quad & \sum_{c=1}^{C} z_c r_c \left(\beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}_c) \right) \\ \text{s.t.} \quad & \underline{I} \le \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \le \overline{I}, \ \|\boldsymbol{\beta}\| \le \hat{\Gamma}_1, \end{aligned} \qquad \begin{aligned} \min_{(\epsilon(\mathbf{x}_c))_{c=1}^{C}:} \quad & \sum_{c=1}^{C} z_c r_c \epsilon(\mathbf{x}_c) \\ \text{s.t.} \quad & \|(\epsilon(\mathbf{x}_c))\|_{\mathsf{res}} \le \hat{\Gamma}_2, \end{aligned} \qquad \begin{aligned} \min_{(v(\mathbf{x}_c))_{c=1}^{C}:} \quad & \sum_{c=1}^{C} z_c r_c v(\mathbf{x}_c) \\ \text{s.t.} \quad & \|(v(\mathbf{x}_c))\|_{\mathsf{link}} \le \hat{\kappa} \end{aligned}$$

Consider the second optimization problem. By the Cauchy-Schwarz inequality, the optimal value is at least $-\hat{\Gamma}_2 \left\| (z_c r_c)_{c=1}^{C} \right\|_{\mathsf{res}}^{*}$. In fact, the optimal value is exactly this quantity by definition of the dual norm.

An entirely analogous argument holds for the third optimization problem, which has optimal value $-\hat{\kappa} \left\| (z_c r_c)_{c=1}^{C} \right\|_{\mathsf{link}}^{*}$.

It remains to evaluate the first optmization problem. For $\hat{\Gamma}_1 > 0$, $\beta_0 = (\overline{I} - \underline{I})/2$, $\boldsymbol{\beta} = \mathbf{0}$, is a strictly feasible solution. It follows that Slater's condition holds, and we have strong duality. Dualizing the two linear inequalities and rearranging yields the Lagrangian dual $\sup_{\gamma_1, \gamma_2 \ge 0} \mathcal{G}(\gamma_1, \gamma_2)$ where

$$\mathcal{G}(\gamma_1, \gamma_2) = \gamma_1 \underline{I} - \gamma_2 \overline{I} \ + \ \min_{\beta_0} \ \beta_0 \sum_{c=1}^{C} (z_c r_c - \gamma_1 + \gamma_2) \ + \ \min_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\| \le \hat{\Gamma}_1} \ \sum_{g=1}^{G} \left(\sum_{c=1}^{C} z_c r_c \phi_g(\mathbf{x}_c) - (\gamma_1 - \gamma_2) \mu_g \right) \beta_g.$$

The first minimization is finite only if $\gamma_1 - \gamma_2 = \sum_{c=1}^{C} z_c r_c$. By the Cauchy-Schwarz inequality the second minimization is equal to

$$\left\| \left(\sum_{c=1}^{C} z_c r_c \phi_g(\mathbf{x}_c) - (\gamma_1 - \gamma_2) \mu_g \right)_{g=1}^{G} \right\|^{*}.$$

Substituting these values above yields the dual problem:

$$\begin{aligned} \sup_{\gamma_1,\gamma_2} \quad & \gamma_1 \underline{I} - \gamma_2 \overline{I} + \left\| \left(\sum_{c=1}^{C} z_c r_c \phi_g(\mathbf{x}_c) - (\gamma_1 - \gamma_2)\mu_g \right)_{g=1}^{G} \right\|^* \\ \text{s.t.} \quad & \gamma_1 - \gamma_2 = \sum_{c=1}^{C} z_c r_c. \end{aligned}$$

Since $\overline{I} > \underline{I}$, we claim at optimality that $\gamma_2 = 0$. Indeed, if this were not true, then the solution $(\gamma_1 - \gamma_2, 0)$ is still feasible, but yields a better objective value than (γ_1, γ_2). We conclude that at optimality, $\gamma_1^* = \sum_{c=1}^{C} z_c r_c$ and $\gamma_2^* = 0$. Substituting above, combining the three subproblems and simplifying proves the result. □

THEOREM EC.1. *Problem* (12) *is* NP-Complete *even if* $G = 1$ *and* $\hat{\kappa} = \hat{\Gamma}_2 = 0$, $\phi(\mathbf{x})$ *is binary-valued and* $\|\cdot\|^*$ *is the* ℓ_p*-norm.*

Proof of Theorem EC.1. We reduce to the well-known *NP-Complete* problem Subset Sum. We first state the decision version of subset sum and the relevant special case of problem (12).

Subset Sum : *Given natural numbers* $a_1, \ldots, a_M$*, and a target number* $T > 0$*, is there a subset of* $N \subseteq \{a_1, ..., a_M\}$ *that adds up to precisely* T*?*

Decision Version of Special Case of Problem (12): *Given parameters* $\underline{I}$, $\hat{\Gamma}_1$, K *and* μ*, sequences* $r_c \geq 0$ *and* $\phi(\mathbf{x}_c) \in \{0, 1]\}$ *for* $c = 1, \ldots, C$*, and a target value* Q*, is the objective value of*

$$\max_{\mathbf{z} \in \mathcal{Z}} \quad \underline{I} \sum_{c=1}^{C} z_c r_c - \hat{\Gamma}_1 \left| \sum_{c=1}^{C} z_c r_c (\phi(\mathbf{x}_c) - \mu) \right|$$

at least Q*?*

We next describe the reduction.: Given an instance of Subset Sum with positive integers $\{a_1, \ldots, a_M\}$ and target sum value T, for any $\underline{I} > 0$, let $\hat{\Gamma}_1 > 2\underline{I}$, and take $C = K = M + 1$, $\mu = 0.5$, $Q = 2T\underline{I}$, and

$$r_c = \begin{cases} a_c & \text{if } c = 1, \ldots, M \\ T & \text{if } c = M + 1 \end{cases}, \quad \phi(\mathbf{x}_c) = \begin{cases} 0 & \text{if } c = 1, \ldots, M \\ 1 & \text{if } c = M + 1 \end{cases}.$$

Let $\mathbf{z}^*$ be the solution to Eq. (12) with these parameters. We will prove that the objective value is at least Q if and only if the answer to Subset Sum is "Yes."

Assume without loss of generality that $T \leq \sum_{c=1}^{m} a_c$, else the answer to the Subset Sum problem is trivially, "No." Thus, we have the simple bound

$$\underline{I} \sum_{c=1}^{C} z_c^* r_c - \hat{\Gamma}_1 \left| \sum_{c=1}^{C} z_c^* r_c (\phi(\mathbf{x}_c) - \mu) \right| \leq \underline{I} \sum_{c=1}^{C} z_c^* r_c \leq 2T\underline{I}. \tag{EC.6}$$

Now suppose that the optimal objective is at least Q. We claim that z^*_{M+1} must equal 1. Indeed, if $z^*_0 = 0$, then the objective can be rewritten as

$$\begin{aligned}
\underline{I}\sum_{c=1}^{C} z_c r_c - \hat{\Gamma}_1 \left|\sum_{c=1}^{C} z_c r_c(\phi(\mathbf{x}_c) - \mu)\right| &= \underline{I}\sum_{c=1}^{M} z^*_c r_c - \hat{\Gamma}_1 \left|\sum_{c=1}^{M} z_c r_c \cdot \frac{-1}{2}\right| \quad \text{(since } \phi(\mathbf{x}_c) = 0 \text{ for all } c = 1,\ldots,M\text{)} \\
&= (\underline{I} - \hat{\Gamma}_1/2)\sum_{c=1}^{M} z^*_c r_c \\
&\leq 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(since } \hat{\Gamma} > 2\underline{I}\text{).}
\end{aligned}$$

This contradicts the assumption that the optimal objective is at least Q.

Thus, if the optimal objective is at least Q, each of the inequalities in Eq. (EC.6) must be equalities. Furthermore, since $z^*_{M+1} = 1$, it must be that $\sum_{c=1}^{M} z^*_c r_c = T$, i.e, these z^* encode the relevant subset and the answer to Subset Sum is "Yes."

Now suppose that the answer to Subset Sum is "Yes." Then, consider a solution $\mathbf{z}_c$ for $c = 1,\ldots,M$ encoded by this subset and $z_{M+1} = 1$. The objective value of this solution is

$$\underline{I}\sum_{c=1}^{C} z_c r_c - \hat{\Gamma}_1 \left|\sum_{c=1}^{C} z_c r_c(\phi(\mathbf{x}_c) - \mu)\right| \;=\; 2T\underline{I} - \hat{\Gamma}_1 \left|T(-.5) + T(1-.5)\right| \;=\; 2T\underline{I} \;=\; Q.$$

Thus, the optimal value must be at least Q. This completes the proof. □

Proof of Corollary 2. Substituting in the appropriate norms to problem (12) yields

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad (\underline{I} - \hat{\Gamma}_2 - \hat{\kappa})\sum_{c=1}^{C} z_c r_c - \hat{\Gamma}_1 \left\| \left(\sum_{c=1}^{C} z_c r_c(\phi_g(\mathbf{x}_c) - \mu_g)\right)_{g=1}^{G} \right\|^*.$$

If $\underline{I} - \hat{\Gamma}_2 - \hat{\kappa} < 0$, then both terms of the objective are non-negative and $\mathbf{z}^* = 0$ is an optimal solution. Else, we can factor out this term from the maximization yielding problem (13). □

Proof of Corollary 3. Define $\mu_{G+1} \equiv 1 - \boldsymbol{\mu}^T\mathbf{e}$, where $\mathbf{e} \in \Re^G$ is a vector of ones. We first claim

$$\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(\boldsymbol{\mu})^{-1} + \mu_{G+1}^{-1}\mathbf{e}\mathbf{e}^T. \tag{EC.7}$$

Indeed, we compute directly

$$\boldsymbol{\Sigma}\cdot\boldsymbol{\Sigma}^{-1} = \mathbf{I} + \mu_{G+1}^{-1}\boldsymbol{\mu}\mathbf{e}^T - \boldsymbol{\mu}\mathbf{e}^T - \mu_{G+1}^{-1}(\boldsymbol{\mu}^T\mathbf{e})\boldsymbol{\mu}\mathbf{e}^T = \mathbf{I} + \left(\mu_{G+1}^{-1} - 1 - \mu_{G+1}^{-1}(1-\mu_{G+1})\right)\boldsymbol{\mu}\mathbf{e}^T = \mathbf{I}$$

and

$$\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{\Sigma} = \mathbf{I} - \mathbf{e}\boldsymbol{\mu}^T + \mu_{G+1}^{-1}\mathbf{e}\boldsymbol{\mu}^T - \mu_{G+1}^{-1}(\boldsymbol{\mu}^T\mathbf{e})\mathbf{e}\boldsymbol{\mu}^T = \mathbf{I} - \left(1 - \mu_{G+1}^{-1} + \mu_{G+1}^{-1}(1-\mu_{G+1})\right)\mathbf{e}\boldsymbol{\mu}^T = \mathbf{I},$$

which proves the claim. Problem (13) is equivalent to

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad \sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \left\| \left(\sum_{c=1}^{C} z_c r_c(\mathbb{I}(\mathbf{x}_c \in \mathcal{X}_g) - \mu_g)\right)_{g=1}^{G} \right\|_{\boldsymbol{\Sigma}^{-1}}. \tag{EC.8}$$

For notational convenience, let us define $f_g \equiv \sum_{c=1}^{C} z_c r_c(\mathbb{I}(\mathbf{x}_c \in \mathcal{X}_g) - \mu_g)$ for all $g = 1, \ldots, G$. Consequently,

$$\begin{aligned}
\mathbf{f}^T\mathbf{e} &= \sum_{c=1}^{C} z_c r_c (\sum_{g=1}^{G} \mathbb{I}(\mathbf{x}_c \in \mathcal{X}_g) - \sum_{g=1}^{G} \mu_g) \\
&= \sum_{c=1}^{C} z_c r_c \left((1 - \mathbb{I}(\mathbf{x}_c \in \mathcal{X}_{G+1})) - (1 - \mu_{G+1})\right) \\
&= \sum_{c=1}^{C} z_c r_c \left(\mu_{G+1} - \mathbb{I}(\mathbf{x}_c \in \mathcal{X}_{G+1})\right).
\end{aligned}$$

Simplifying (EC.8) yields:

$$\begin{aligned}
\text{Eq. (EC.8)} &= \sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \sqrt{\mathbf{f}^T \boldsymbol{\Sigma}^{-1} \mathbf{f}} \\
&= \sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \sqrt{\mathbf{f}^T \left(\operatorname{diag}(\boldsymbol{\mu})^{-1} + \mu_{G+1}^{-1} \mathbf{e}\mathbf{e}^T\right) \mathbf{f}} \\
&= \sum_{c=1}^{C} z_c r_c - \frac{\hat{\Gamma}}{\underline{I} - \hat{\Gamma}_2 - \hat{\kappa}} \sqrt{\mathbf{f}^T \operatorname{diag}(\boldsymbol{\mu})^{-1} \mathbf{f} + (\mathbf{f}^T\mathbf{e})^2 \mu_{G+1}^{-1}},
\end{aligned}$$

which yields (15). Thus, the proof is complete. □

Proof of Corollary 4. The proof follows directly by applying Corollary 2 and by definition of dual norm. □

Proof of Corollary 5 To prove the first inequality, note that there exists $\boldsymbol{\Gamma}, \kappa$ such that the true CATE $\Psi^*(\cdot) \equiv \mathbb{E}[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c]$ is a member of $\mathcal{U}_{\boldsymbol{\Gamma},\kappa}$. So we have

$$\sum_{c=1}^{C} r_c z_c^{\text{Rob}} \mathbb{E}[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c] \geq \min_{\Psi_{\mathsf{C}}^*(\cdot) \in \mathcal{U}_{\boldsymbol{\Gamma},\kappa}} \sum_{c=1}^{C} r_c z_c^{\text{Rob}} \Psi^*(\mathbf{x}_c), \tag{EC.9}$$

and applying Theorem 3 yields the result. To see the second equality, notice that for any candidate CATE $\Psi_{\mathsf{C}}(\cdot) \in \mathcal{U}_{\hat{\boldsymbol{\Gamma}},\hat{\kappa}}$,

$$\min_{\Psi_{\mathsf{C}}(\cdot) \in \mathcal{U}_{\hat{\boldsymbol{\Gamma}},\hat{\kappa}}} \sum_{c=1}^{C} r_c z_c^{\text{Rob}} \Psi_{\mathsf{C}}(\mathbf{x}_c) \geq 0,$$

where the inequality follows because $z = 0$ is a feasible solution to the robust problem while $\mathbf{z}^{\text{Rob}}$ is an optimal solution. Continuing Eq. (EC.9), we have

$$\sum_{c=1}^{C} r_c z_c^{\text{Rob}} \mathbb{E}[\delta_{\tilde{c}} \mid \mathbf{x}_{\tilde{c}} = \mathbf{x}_c] \geq \min_{\Psi_{\mathsf{C}}^*(\cdot) \in \mathcal{U}_{\boldsymbol{\Gamma},\kappa}} \sum_{c=1}^{C} r_c z_c^{\text{Rob}} \Psi^*(\mathbf{x}_c) - \min_{\Psi_{\mathsf{C}}(\cdot) \in \mathcal{U}_{\hat{\boldsymbol{\Gamma}},\hat{\kappa}}} \sum_{c=1}^{C} r_c z_c^{\text{Rob}} \Psi_{\mathsf{C}}(\mathbf{x}_c).$$

Applying Theorem 3 again yields the result, which completes the proof. □

To show that problem (17) can be solved by bisection search for any α, it suffices to show that $\mathbf{z}(0)$ equals reward scoring, $\lim_{\lambda \to \infty} \mathbf{z}(\lambda) = 0$ and that the constraint is monotonic in λ. The first two claims are clear from the definition of $\mathbf{z}(\lambda)$. We prove the last:

THEOREM EC.2 **(Monotonicity of $\mathbf{z}(\lambda)$).** *The function $\lambda \to \sum_{c=1}^{C} r_c z_c(\lambda)$ is non-increasing.*

Proof of Theorem EC.2. Let $0 \le \lambda_1 < \lambda_2$. Then, from optimality of $\mathbf{z}(\lambda_1)$,

$$\sum_{c=1}^{C} r_c z_C(\lambda_1) - \lambda_1 \left\| \left(\sum_{c=1}^{C} z_c(\lambda_1) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* \ge \sum_{c=1}^{C} r_c z_C(\lambda_2) - \lambda_1 \left\| \left(\sum_{c=1}^{C} z_c(\lambda_2) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*. \tag{EC.10}$$

Similarly, from the optimality of $\mathbf{z}(\lambda_2)$,

$$\sum_{c=1}^{C} r_c z_C(\lambda_2) - \lambda_2 \left\| \left(\sum_{c=1}^{C} z_c(\lambda_2) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* \ge \sum_{c=1}^{C} r_c z_c(\lambda_1) - \lambda_2 \left\| \left(\sum_{c=1}^{C} z_c(\lambda_1) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*.$$

Adding these two equations and rearranging yields:

$$0 \ge (\lambda_1 - \lambda_2) \left(\left\| \left(\sum_{c=1}^{C} z_c(\lambda_1) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* - \left\| \left(\sum_{c=1}^{C} z_c(\lambda_2) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* \right)$$

Since $\lambda_1 < \lambda_2$, this implies that

$$\left\| \left(\sum_{c=1}^{C} z_c(\lambda_1) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* \ge \left\| \left(\sum_{c=1}^{C} z_c(\lambda_2) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*.$$

Substituting back into Eq. (EC.10) and rearranging shows

$$\begin{aligned} \sum_{c=1}^{C} r_c z_c(\lambda_1) &\ge \sum_{c=1}^{C} r_c z_c(\lambda_2) + \lambda_1 \left(\left\| \left(\sum_{c=1}^{C} z_c(\lambda_1) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* - \left\| \left(\sum_{c=1}^{C} z_c(\lambda_2) r_c(\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^* \right) \\ &\ge \sum_{c=1}^{C} r_c z_c(\lambda_2), \end{aligned}$$

which completes the proof. □

EC.2. Extensions of the Base Model

EC.2.1. Fairness Constraints, and Domain-Specific Knowledge

Adding constraints on $\mathbf{z}$ does not significantly increase the complexity of the robust model. Such constraints can be used to enforce fairness, e.g., that equal numbers of men and women be targeted for treatment. Similarly, it is straightforward to adjust the budget to the form $\mathbf{d}^T \mathbf{z} \le K$ to model the case when different patients have different costs of treatment d_c.

Similarly, adding convex constraints to our uncertainty set (11) does not significantly increase the complexity of the model. Thus, we might incorporate domain-specific knowledge of the structure on $\Psi_{\mathsf{C}}(\cdot)$ by enforcing, e.g., $\beta_g = 0$ for some g, or by bounding its magnitude, as in $l_c \le \Psi_{\mathsf{C}}(\mathbf{x}_c) \le u_c$. Applying standard techniques yields a corresponding robust counterpart.

EC.2.2. Incorporating Evidence from Multiple Studies

When there are multiple, distinct studies providing evidence that the treatment is effective, we would prefer to incorporate all of them into our model. For concreteness, consider the case of J studies with corresponding confidence intervals $[\underline{I}^j, \overline{I}^j]$, description functions $\phi_g^j(\cdot)$ and statistics μ_g^j for all $g = 1, \ldots, G_j$ and $j = 1, \ldots, J$. With these data, let

$$\mathcal{U}^j_{\hat{\mathbf{\Gamma}}^j, \hat{\kappa}^j} = \Bigg\{ \Psi_{\mathsf{C}}(\cdot) : \mathcal{X} \mapsto \mathbb{R} \;\Bigg|\; \exists \epsilon(\cdot) : \mathcal{X} \mapsto \mathbb{R},\ \beta_0 \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^{G_j},\ \text{s.t.}\ \ \Psi^{\mathsf{S}}(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G_j} \beta_g \phi_g^j(\mathbf{x}) + \epsilon(\mathbf{x}), \quad \text{(EC.11)}$$
$$\underline{I}^j \le \beta_0 + \sum_{g=1}^{G_j} \beta_g \mu_g^j \le \overline{I}^j,\ \ \left\| \left(\Psi^{\mathsf{S}}(\mathbf{x}_c) - \Psi_{\mathsf{C}}(\mathbf{x}_c)\right)_{c=1}^{C} \right\|_{\mathsf{link}} \le \hat{\kappa}^j,\ \ \|\boldsymbol{\beta}\| \le \hat{\Gamma}_1^j,\ \ \left\| (\epsilon(\mathbf{x}_c))_{c=1}^{C} \right\|_{\mathsf{res}} \le \hat{\Gamma}_2^j, \Bigg\}$$

be our usual uncertainty set built using the study evidence of the j^{th} paper. Then a natural extension of our robust model is

$$\max_{\mathbf{z} \in \mathcal{Z}} \ \min_{\Psi_{\mathsf{C}}(\cdot) \in \bigcap_{j=1}^J \mathcal{U}^j_{\hat{\mathbf{\Gamma}}^j, \hat{\kappa}^j}} \ \sum_{c=1}^{C} z_c r_c \Psi_{\mathsf{C}}(\mathbf{x}_c), \tag{EC.12}$$

i.e., to consider worst-case performance over models for the CATE which are consistent with each of the studies. Again, using fairly standard techniques, we can form the robust counterpart:

THEOREM EC.3 **(Robust Counterpart for Multiple Papers)**. *Problem* (EC.12) *is equivalent to*

$$\begin{array}{ll} \max\limits_{\mathbf{z}, \mathbf{w}} & \displaystyle\sum_{j=1}^{J} \left(\underline{I}^j \sum_{c=1}^{C} w_c^j - \hat{\Gamma}_1^j \left\| \left(\sum_{c=1}^{C} w_c^j (\phi_g^j(\mathbf{x}_c) - \mu_g^j) \right)_{g=1}^{G_j} \right\|^* - \hat{\Gamma}_2^j \left\| \left(w_c^j\right)_{c=1}^{C} \right\|_{\mathsf{res}}^* - \hat{\kappa}^j \left\| \left(w_c^j\right)_{c=1}^{C} \right\|_{\mathsf{link}}^* \right) \\ s.t. & \displaystyle\sum_{j=1}^{J} w_c^j = z_c r_c \quad c = 1, \ldots C. \end{array}$$

Proof. Let $\mathcal{U}^j \subseteq \mathbb{R}^C$ be the set $\left\{ (\Psi_{\mathsf{C}}(\mathbf{x}_c)\ : c = 1, \ldots, C) \mid \Psi_{\mathsf{C}} \in \mathcal{U}^j_{\hat{\mathbf{\Gamma}}_j, \hat{\kappa}_j} \right\}$. In words, $\mathcal{U}^j$ are the set of possible realizations of the candidate CATE on the candidate population that are consistent with the study evidence from the j^{th} paper. Then, for a fixed $\mathbf{z}$, the inner minimization of Problem (EC.12) is equivalent to

$$-\max_{\mathbf{p}} \quad \sum_{c=1}^{C} (-z_c r_c) p_c \qquad \text{s.t.} \quad \mathbf{p} \in \bigcap_{j=1}^{J} \mathcal{U}^j.$$

We recognize the maximum as the support function of the set $\bigcap_{j=1}^J \mathcal{U}^j$ evaluated at $(-z_c r_c : c = 1 \ldots C)$. Standard results allow us to re-express this support function in terms of the support functions of $\mathcal{U}^j$ (see, e.g., Ben-Tal et al. (2015)). Specifically, the above optimization is equivalent to

$$-\min_{\mathbf{w}} \quad \sum_{j=1}^{J} \delta^*(\mathbf{w}^j \mid \mathcal{U}^j) \qquad \text{s.t.} \quad \sum_{j=1}^{J} \mathbf{w}_c^j = -z_c r_c, \quad c = 1, \ldots, C,$$

where $(\delta^*\mathbf{w} \,|\, \mathcal{U}^j) \equiv \sup_{\mathbf{p}\in\mathcal{U}^j} \mathbf{p}^T\mathbf{w}$ is the support function of $\mathcal{U}^j$. Pass the negative sign through the minimization and make the transformation $\mathbf{w}^j \to -\mathbf{w}^j$ to write

$$\max_{\mathbf{w}} \quad \sum_{j=1}^{J} -\delta^*(-\mathbf{w}^j \,|\, \mathcal{U}^j) \qquad \text{s.t.} \quad \sum_{j=1}^{J} \mathbf{w}_c^j = z_c r_c, \quad c=1,\ldots,C.$$

Finally, note that $-\delta^*(-\mathbf{w}^j \,|\, \mathcal{U}^j) = \min_{\mathbf{p}\in\mathcal{U}^j} \sum_{c=1}^{C} w_c^j p_c$, and this minimization is precisely the inner problem in the proof of Theorem 3. Applying that result and simplifying completes the proof. □

In principle, one could specify the each of the norms and parameters $\hat{\Gamma}_1^j, \hat{\Gamma}_2^j, \hat{\kappa}^j$, separately, although practical considerations would favor taking them to be equal. Importantly, the theorem emphasizes that in our meta-analysis setting, adding additional evidence does not affect our uncertainty set by simply shrinking its radius as in more traditional data-driven robust optimization models. Rather, additional evidence adds additional constraints to the set, which yields a more complex robust counterpart.

EC.2.3. Modeling Uncertainty in r_c

In many applications we are not interested in (dollar) rewards, but rather aggregate benefit to patients, in which case setting $r_c = 1$ is a natural choice. Even in settings such as our case-study, where one *is* interested in monetary savings, there is often detailed covariate information available for the candidate population, e.g., medical history and past ED visits, which can be used to build high-quality estimates $\hat{\mathbf{r}}$ of these savings from historical data. In these cases, the uncertainty in r_c is often relatively small, much smaller than the uncertainty in δ_c, and approximating $r_c \approx \hat{r}_c$ is reasonable.

That said, from a theoretical point of view, one could imagine settings where the uncertainty in $\mathbf{r}$ is large, and one wishes to"robustify" this parameter. As a simple example, suppose we model $\mathbf{r} \in \mathcal{U}_{\Gamma_r} \equiv \{\hat{\mathbf{r}} + \Delta\mathbf{r} : \|\boldsymbol{\Delta}\mathbf{r}\|_{\mathsf{r}} \le \Gamma_r\}$ for some point estimate $\hat{\mathbf{r}}$ bounded error $\Delta\mathbf{r}$, and we wish to solve

$$\max_{\mathbf{z}\in\mathcal{Z}} \min_{\Psi(\cdot)\in\mathcal{U}_{\hat{\mathbf{\Gamma}},\hat{\kappa}}} \min_{r\in\mathcal{U}_{\Gamma_r}} \sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c).$$

A straightforward computation shows this problem is equivalent to

$$\max_{\mathbf{z}\in\mathcal{Z}} \min_{\Psi(\cdot)\in\mathcal{U}_{\hat{\mathbf{\Gamma}},\hat{\kappa}}} \sum_{c=1}^{C} z_c \hat{r}_c \Psi(\mathbf{x}_c) - \Gamma_r \left\| (z_c\Psi(\mathbf{x}_c))_{c=1}^{C} \right\|_{\mathsf{r}}^*$$

where $\|\cdot\|_{\mathsf{r}}$ and $\|\cdot\|_{\mathsf{r}}^*$ are dual norms.

We recognize this as a robust problem where the uncertainty occurs in a *convex* fashion in the inner problem. There are a variety of approaches to attacking such problems when $\mathcal{U}_{\hat{\mathbf{\Gamma}}}$ is polyhedral, including, e.g., vertex enumeration and converting the problem to adjustable linear program (Zhen et al. 2017a, Den Hertog 2018) which can be solved exactly via Fourier-Motzkin elimination (Zhen

et al. 2017b) or approximately via decision-rules (Kuhn et al. 2011). For clarity, $\mathcal{U}_{\hat{\mathbf{\Gamma}},\hat{\kappa}}$ will be polyhedral whenever the norms defining it are (weighted) ℓ_1 or ℓ_∞ norms.

Solving robust problems with convex uncertainty can be computationally demanding. Each of the above approaches offers its own strengths and drawbacks. The best approach is often application dependent. Since our target application is well-modeled by a known r_c, we leave a comprehensive study of the computational merits of each of the above approaches to future work.

EC.2.4. Other Forms of Covariate Matching as Regularization

Section 4.5 showed that with appropriately chosen norms in our uncertainty set, we can recover several well-known covariate matching techniques as regularizers in the robust counterpart. In this section we show how, given a general covariate matching technique, one can modify the construction of Eq. (11) to obtain the corresponding uncertainty set.

Given a candidate targeting $\mathbf{z}$, let $\mathbf{r}\circ\mathbf{z} = (r_c z_c)_{c=1}^C$, and let $\mathbf{w}(\mathbf{z}) = \frac{\mathbf{r}\circ\mathbf{z}}{\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z})} \in \mathbb{R}_+^C$. We can interpret $\mathbf{w}(\mathbf{z})$ as a discrete probability distribution on $\mathcal{X}$ which assigns mass $w_c(\mathbf{z})$ to each point $\mathbf{x}_c \in \mathcal{X}$, $c = 1,\ldots,C$. We will commit a small abuse of notation and refer to $\mathbf{w}(\mathbf{z})$ and this probability distribution interchangeably. Similarly, we let $\mathbb{P}^{\mathsf{S}}$ denote the empirical distribution of the covariates on $\mathcal{X}$ in the study population, i.e., the discrete probability distribution that assigns mass $1/S$ to each point $\mathbf{x}^s \in \mathcal{X}$, $s = 1,\ldots,S$.

Intuitively, covariate matching techniques seek a group such that the distribution of covariates in this group closely matches that of some other, fixed group of interest. (In causal studies, this often amounts to finding a control group who closely matches the treatment group.) We restrict attention to covariate matching techniques of the form $\min_{\mathbf{z}} d(\mathbf{w}(\mathbf{z}), \mathbb{P}^{\mathsf{S}})$, where $d(\cdot,\cdot)$ is a function measuring the "distance" between two probability distributions defined on $\mathcal{X}$. (This is where we commit our aforementioned abuse of notation.) We write "distance" in quotation marks because we do not require that the function be a metric (see below for examples).

Almost all common covariate matching techniques can be written in this form for some $d(\cdot,\cdot)$. For example, we might take $d(\mathbf{w}(\mathbf{z}),\ \mathbb{P}^{\mathsf{S}})$ to be an integral probability metric, such as

$$\sup_{A\subseteq\mathcal{X}} \left| \sum_{c=1}^C w_c(\mathbf{z}) \mathbb{I}(\mathbf{x}_c \in A) - \frac{1}{S}\sum_{s=1}^S \mathbb{I}(\mathbf{x}^s \in A) \right| \qquad \text{(Total Variation Distance)}$$

or

$$\sup_{\substack{f:\mathcal{X}\to\mathbb{R},\\ f \text{ is 1-Lipschitz}}} \left| \sum_{c=1}^C w_c(\mathbf{z}) f(\mathbf{x}_c) - \frac{1}{S}\sum_{s=1}^S f(\mathbf{x}^s) \right| \qquad \text{(Wasserstein Distance).}$$

These metrics minimize the total variation and Wasserstein distance between the distribution of covariates in the target population and study population, respectively. Alternatively, we can take $d(\mathbf{w}(\mathbf{z}),\ \mathbb{P}^{\mathsf{S}})$ to be a general ϕ-divergence metric,

$$\frac{1}{S}\sum_{s=1}^{S}\phi\left(\frac{\sum_{c=1}^{C} w_c(\mathbf{z})\mathbb{I}(\mathbf{x}_c=\mathbf{x}^s)}{\frac{1}{S}\sum_{r=1}^{S}\mathbb{I}(\mathbf{x}^r=\mathbf{x}^s)}\right) \qquad (\phi\text{-divergence}),$$

where $\phi(\cdot)$ is a convex function satisfying $\phi(1)=0,\ 0\phi(a/0)\equiv a\lim_{t\to\infty}\phi(t)/t$ for $a>0$ and $0\phi(0/0)\equiv 0$.[6] By specializing the function ϕ, ϕ-divergences recover many well-known probability metrics including relative entropy, Hellinger-distance, and the Cressie-Read divergences. See Ben-Tal et al. (2013) for more examples.

Importantly, our earlier covariate matching results, namely, Corollaries 3 and 4, can also be obtained as special cases of this framework. Loosely speaking, we take $d(\cdot,\cdot)$ to be the function of the two probability distributions which first maps each distribution to the expected value of the description functions with respect to that distribution, and then applies a function to these expected values.

More specifically, consider the function $d(\mathbf{w}(\mathbf{z}),\mathbb{P}^{\mathsf{S}})=\left\|\sum_{c=1}^{C} w_c(\mathbf{z})\mathbf{x}_c-\frac{1}{S}\sum_{s=1}^{S}\mathbf{x}^s\right\|_{\mathbf{V}^{-1}}$, which effectively computes the mean of each distribution and then computes the weighted ℓ_2 norm between the resulting means. Minimizing this $d(\cdot,\cdot)$ is equivalent to Mahanoblis matching (compare to Corollary 4).

Similarly, suppose as in Corollary 3, that there exists a partition $\mathcal{X}=\bigcup_{g=1}^{G+1}$ and the G description functions are given by $\mathbb{I}(\mathbf{x}\in\mathcal{X}_g)$. Then the function $d(\mathbf{w}(\mathbf{z}),\mathbb{P}^{\mathsf{S}})$ given by

$$\sqrt{\sum_{g=1}^{G+1}\frac{(q_{\mathbf{z},g}-\mu_g)^2}{\mu_g}},$$

where

$$q_{\mathbf{z},g}=\sum_{c=1}^{C} w_c(\mathbf{z})\mathbb{I}(\mathbf{x}_c\in\mathcal{X}_g) \quad \text{and} \quad \mu_g=\frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(\mathbf{x}^s\in\mathcal{X}_g)$$

effectively computes the mean of each description function under the two measures, and then computes the χ^2-distance between them. Minimizing this $d(\cdot,\cdot)$ is equivalent to χ^2 matching as in Corollary 3.

We stress that in contrast to integral probability metrics and general ϕ-divergences, the last two examples only depend on $\mathbb{P}^{\mathsf{S}}$ via the statistics of the description functions, not the entire distribution.

[6] The use of ϕ in defining the ϕ-divergences is cannonical and unfortunately conflicts with our use of ϕ in defining the description functions. We will only refer to ϕ-divergences here, and stress, that with this exception of this appendix, all references to ϕ refer to description functions in the study evidence.

In summary, the possibilities for distance functions and covariate matching techniques under this framework are numerous and our list is non-exhaustive. We refer the reader to Gibbs and Su (2002), Pardo (2005) and references therein for further examples and discussion.

Given such a distance function $d(\cdot,\cdot)$, we can define a modification of Eq. (11) that yields this distance function as a regularizer, i.e., we can obtain a "covariate matching as regularization" interpretation of the robust counterpart for a general covariate matching technique. We will require that d satisfy some mild conditions. Each of the examples above satisfies these conditions.

ASSUMPTION EC.1 **(Regularity Conditions of Distance Function)**. *We assume that*

1. $d(\cdot,\mathbb{P}^{\mathsf{S}})$ *is convex and lower-semicontinuous in its first argument,*
2. $d(\mathbb{P}^{\mathsf{S}},\mathbb{P}^{\mathsf{S}})=0$, *and*
3. $d(\mathbb{P},\mathbb{P}^{\mathsf{S}})>-\infty$ *for all* $\mathbb{P}$.

THEOREM EC.4 **(General Covariate Matching as Regularization)**. *Suppose* $d(\cdot,\cdot)$ *satisfies Assumption EC.1. Let*

$$\mathcal{U}_{\hat{\mathbf{\Gamma}},\hat{\kappa},d}=\left\{\Psi_{\mathsf{C}}(\cdot):\mathcal{X}\mapsto\mathbb{R}\;\middle|\;\begin{array}{l}\exists\epsilon(\cdot):\mathcal{X}\mapsto\mathbb{R},\;\Psi^{\mathsf{S}}(\cdot):\mathcal{X}\mapsto\mathbb{R},\quad\theta,I\in\mathbb{R},\quad\mathbf{y}\in\mathbb{R}^C,\;\;s.t.\\ \Psi^{\mathsf{S}}(\mathbf{x}_c)=I+\theta-y_c+\epsilon(\mathbf{x}_c),\quad c=1,\ldots,C,\\ \underline{I}\le I\le\overline{I},\quad\left\|\left(\Psi^{\mathsf{S}}(\mathbf{x}_c)-\Psi_{\mathsf{C}}(\mathbf{x}_c)\right)_{c=1}^{C}\right\|_{\mathsf{link}}\le\hat{\kappa},\quad d^*\left(\frac{\mathbf{y}}{\hat{\Gamma}_1}\right)\le\frac{\theta}{\hat{\Gamma}_1},\quad\left\|(\epsilon(\mathbf{x}_c))_{c=1}^{C}\right\|_{\mathsf{res}}\le\hat{\Gamma}_2,\end{array}\right\}, \tag{EC.13}$$

where $d^*(\mathbf{y})\equiv\sup_{\mathbf{w}}\mathbf{y}^\top\mathbf{w}-d(\mathbf{w},\mathbb{P}^{\mathsf{S}})$ *is the convex conjugate of* $d(\cdot,\mathbb{P}^{\mathsf{S}})$. *Then, the robust targeting problem Eq.* (8) *with uncertainty set Eq.* (EC.13) *is equivalent to*

$$\max_{\mathbf{z}\in\mathcal{Z}}\quad\underline{I}\sum_{c=1}^{C}z_cr_c-\hat{\Gamma}_1\left(\sum_{c=1}^{C}r_cz_c\right)\cdot d\left(\mathbf{w}(\mathbf{z}),\mathbb{P}^{\mathsf{S}}\right)-\hat{\Gamma}_2\left\|(z_cr_c)_{c=1}^{C}\right\|_{\mathsf{res}}^{*}-\hat{\kappa}\left\|(z_cr_c)_{c=1}^{C}\right\|_{\mathsf{link}}^{*}. \tag{EC.14}$$

Proof. The proof follows the proof of Theorem 3 closely. Indeed, for a fixed $\mathbf{z}$, we can write $\Psi_C(\mathbf{x}_c)=I+\theta-y_c+\epsilon(\mathbf{x}_c)+\Psi_C(\mathbf{x}_c)-\Psi^{\mathsf{S}}(\mathbf{x}_c)$. With this substitution, the inner minimization similarly decouples into the sum of four minimization problems:

$$\begin{array}{lll}\min\limits_{I} & I\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z}) & \min\limits_{\theta,\mathbf{y}} \quad (\theta\mathbf{e}-\mathbf{y})^\top(\mathbf{r}\circ\mathbf{z})\\ \text{s.t.} & I\in[\underline{I},\overline{I}], & \text{s.t.} \quad d^*\left(\frac{\mathbf{y}}{\hat{\Gamma}_1}\right)\le\frac{\theta}{\hat{\Gamma}_1},\\ \min\limits_{(\epsilon(\mathbf{x}_c))_{c=1}^{C}} & \sum\limits_{c=1}^{C}z_cr_c\epsilon(\mathbf{x}_c) & \min\limits_{(v(\mathbf{x}_c))_{c=1}^{C}} \quad \sum\limits_{c=1}^{C}z_cr_cv(\mathbf{x}_c)\\ \text{s.t.} & \|(\epsilon(\mathbf{x}_c))\|_{\mathsf{res}}\le\hat{\Gamma}_2, & \text{s.t.} \quad \|(v(\mathbf{x}_c))\|_{\mathsf{link}}\le\hat{\kappa}\end{array},$$

where $v(\mathbf{x}_c)$ represents $\Psi^{\mathsf{S}}(\mathbf{x}_c)-\Psi_C(\mathbf{x}_c)$. The solution to the first minimization problem is trivially $I=\underline{I}$. The third and fourth minimization can again be solved using the Cauchy-Schwarz inequality, yielding optimal objectives $\hat{\Gamma}_2\|\mathbf{r}\circ\mathbf{z}\|_{\mathsf{res}}^*$ and $\hat{\kappa}\|\mathbf{r}\circ\mathbf{z}\|_{\mathsf{link}}^*$.

Only the second optimization problem remains. By Lagrange duality, this minimization is equivalent to

$$\sup_{t\geq 0} \quad \left\{ \min_{\theta} \theta(\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z}) - \frac{t}{\hat{\Gamma}_1}) + \min_{\mathbf{y}} \quad -\mathbf{y}^\top(\mathbf{r}\circ\mathbf{z}) + t\cdot d^*\left(\frac{\mathbf{y}}{\hat{\Gamma}_1}\right)\right\}$$

The minimization over θ is finite only if $t = \hat{\Gamma}_1\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z})$. The minimization over $\mathbf{y}$ is equivalent to

$$-t\max_{\mathbf{y}} \quad \hat{\Gamma}_1\mathbf{y}^\top(\mathbf{r}\circ\mathbf{z}/t) - d^*(\mathbf{y}) \quad = \quad -t\cdot d\left(\frac{\hat{\Gamma}_1}{t}\cdot\mathbf{r}\circ\mathbf{z},\ \mathbb{P}^{\mathsf{S}}\right),$$

where we've used Assumption EC.1 to conclude the conjugate of d^* is d itself. Combining and simplifying shows

$$\begin{array}{ll} \min\limits_{\theta,\mathbf{y}} & (\theta\mathbf{e}-\mathbf{y})^\top(\mathbf{r}\circ\mathbf{z}) \\ \text{s.t.} & d^*\left(\dfrac{\mathbf{y}}{\hat{\Gamma}_1}\right) \leq \dfrac{\theta}{\hat{\Gamma}_1}, \end{array} \quad = \quad -\hat{\Gamma}_1\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z})\cdot d\left(\frac{\mathbf{r}\circ\mathbf{z}}{\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z})},\ \mathbb{P}^{\mathsf{S}}\right).$$

Combining the optimal values of all four subproblems proves the theorem. □

Equation (EC.14) decomposes the objective into a portion that maximizes effectiveness under a worst-case homogeneous effect scenario, and three penalties, the first of which is the covariate-matching distance and the second two are as in Theorem 3. In the special case that $\|\cdot\|_{\mathsf{res}} = \|\cdot\|_{\mathsf{link}} = \|\cdot\|_\infty$, we can also simplify the robust counterpart along the lines of Corollary 2, leaving only the covariate-matching distance as a regularizer.

Theorem EC.4 generalizes Theorem 3. Indeed, by choosing $d(\cdot,\cdot)$ appropriately we can recover Eq. (12). Moreover, as already noted above, we can also recover the covariate matching regularizers in Corollaries 3 and 4. Perhaps more importantly, Theorem EC.4 gives an explicit uncertainty set which recovers general covariate matching techniques, e.g., based on total-variation or ϕ-divegences. (As an aside, for most covariate matching techniques listed above, the conjugate d^* is known.) This result thus expands the scope of our "covariate matching as regularization" interpretation of our method.

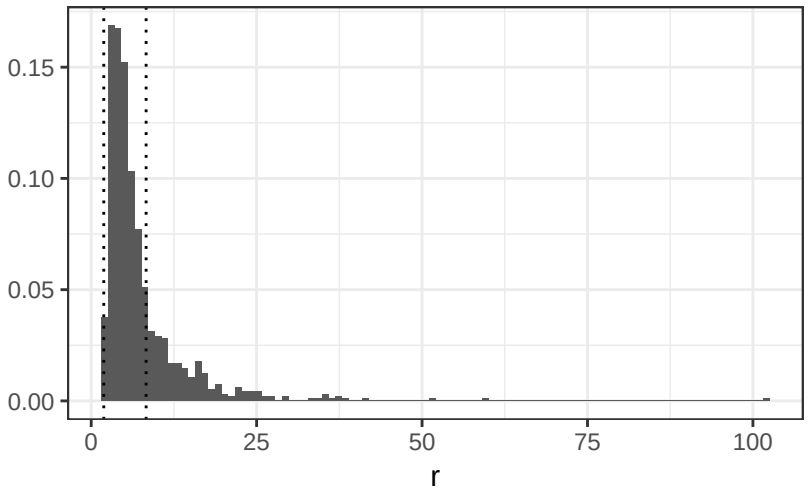
Remark EC.1 (Computational Complexity). Since $d(\cdot,\mathbb{P}^{\mathsf{S}})$ is convex in its first argument, the function $\mathbf{z}\mapsto \mathbf{e}^\top(\mathbf{r}\circ\mathbf{z})\cdot d\left(\frac{\mathbf{r}\circ\mathbf{z}}{\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z})},\ \mathbb{P}^{\mathsf{S}}\right)$ is also convex in $\mathbf{z}$. (Namely, the function $(t,\mathbf{z})\mapsto t\cdot d\left(\frac{\mathbf{r}\circ\mathbf{z}}{t},\ \mathbb{P}^{\mathsf{S}}\right)$ is convex in $(t,\mathbf{z})$ for $t>0$ since it is the perspective function of d, and our desired function is obtained by composing this function with the linear mapping $\mathbf{z}\mapsto(\mathbf{e}^\top(\mathbf{r}\circ\mathbf{z}),\mathbf{z})$.) Thus, Eq. (EC.14) is a mixed-binary convex optimization. Developing algorithms for general mixed-binary convex optimization problems remains an active area of research, but, in our opinion, it is fair to say that from a practical perspective, solving such problems is considerably more difficult than solving mixed-binary linear or mixed-binary convex quadratic optimization problems, and there are many fewer commercial codes available. This increased computational burden makes this approach somewhat less appealing practically than our previous formulations.

REMARK EC.2 (DEPENDENCE ON $\mathbb{P}^{\mathsf{S}}$). As stated earlier, covariate matching techniques based on integral probability metrics or ϕ-divergences typically require access to the full distribution $\mathbb{P}^{\mathsf{S}}$, or, equivalently, $\{\mathbf{x}^s : s = 1, \ldots, S\}$, in order to evaluate $d(\mathbf{w}(\mathbf{z}), \mathbb{P}^{\mathsf{S}})$. Most studies do not report these data, making these types of covariate matching impractical in our application setting. For this reason, we consider Theorem EC.4 to be primarily of theoretical interest. Practical implementations will necessarily have to restrict to covariate matching techniques that only depend on $\mathbb{P}^{\mathsf{S}}$ through the description functions and their statistics.

EC.3. Additional Graphs and Numerical Results

EC.3.1. Graphs from Section 2.3.

Figure EC.1 Histogram of Avg. ED Visit Charges By Patient



Note. ED visit charges are highly concentrated with a long tail. Approximately 75% of charges are between 2 and 8.3 in anonymized monetary units. For comparison, r_c ranges between 0 and 110 in anonymized monetary units, so that approximately 75% of charges occur over 6% of the range.

EC.3.2. Graphs from Section 3

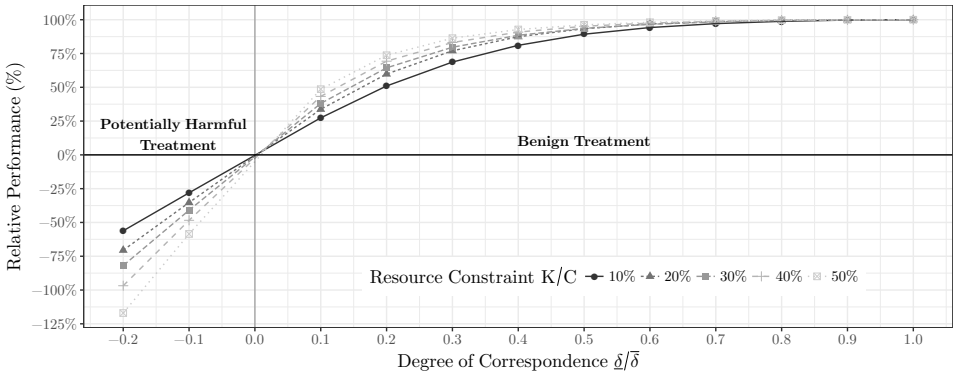
Figure EC.2 shows the worst-case relative performance of outcome scoring (i.e., $ry(0)$-scoring) based on $y_c(0)$. When the case management is benign, outcome scoring performs well, and its performance improves as $\underline{\delta}/\overline{\delta}$ approaches 1. When the resource constraint is relaxed (i.e., as K/C increases), the benefit of outcome scoring improves slightly. To the contrary, when case management could increase the number of ED visits to a smaller degree compared to what it can reduce (i.e., $\underline{\delta}/\overline{\delta} = -0.1$), outcome scoring performs quite badly. Again, this is because the distribution of outcomes $r_c y_c(0)$ has a long tail for a fixed degree of correspondence.

EC.3.3. Graphs from Section 5.3

Recall that for very large values of the Adj. CV parameter, the Robust-2 and Robust-Full-Linear methods may not fully utilize the budget. However, for our dataset, if one specifies the Adj. CV parameter via our method in Section 4.7, both methods fully utilize the budget so long as one specifies $\alpha < 50\%$, i.e., requiring the robust methods achieve at least 50% of the rewards in the nominal case. See Figure EC.3

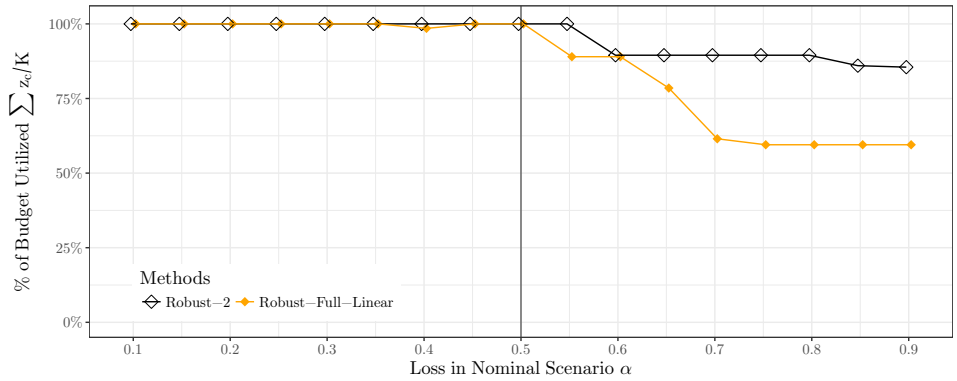
We present the box-plot of rewards for each stratum given by Shumway et al. (2008) in Figure EC.4.

Figure EC.2 Worst-Case Relative Performance of Outcome Scoring ($ry(0)$-Scoring)



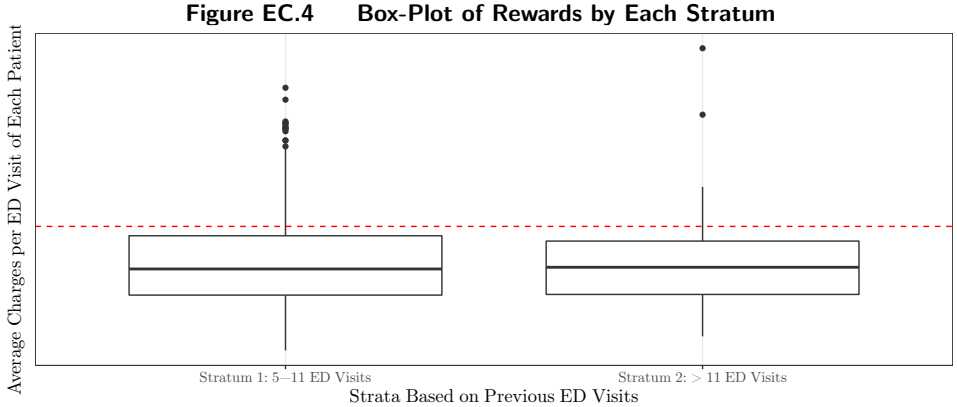
Note. When the treatment is benign, we plot the worst-case relative performance bound (2) provided in Theorem 1. When the treatment is potentially harmful, the worst-case relative performance is $-\infty$, as mentioned in Remark 1. Thus, we plot $\underline{\delta}\sum_{c=1}^{K} r_c y_c(0) e/\overline{\delta}\sum_{c=K+1}^{2K} r_c y_c(0)$ for comparison.

Figure EC.3 Budget Utilization when varying α



Note. For each value of α, we specify the Adj. CV parameter using the method of Section 4.7 and plot the corresponding percentage of the budget utilized.

Figure EC.4 Box-Plot of Rewards by Each Stratum



Note. Each point represents a patient and the monetary values of rewards are anonymized. The dashed line is the cut-off of reward scoring for $K = 200$. Every patient above the line will be targeted by reward scoring.

EC.3.4. Graphs from Section 5.5

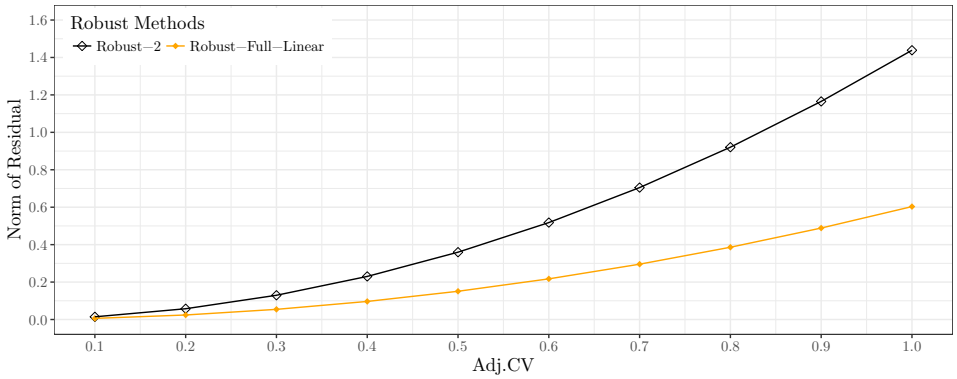
In this section, we leverage Corollary 5 to explain the poor performance of Robust-2 in our experiment in Section 5.5. Recall, given a robust solution $\mathbf{z}^{\text{Rob}}$, the worst-case performance bound in Corollary 5 is

$$(\underline{I} - \Gamma_2^{\text{Rob}} - \kappa^{\text{Rob}}) \sum_{c=1}^{C} z_c^{\text{Rob}} r_c - \Gamma_1^{\text{Rob}} \left\| \left(\sum_{c=1}^{C} z_c^{\text{Rob}} r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|^*,$$

where $\Gamma_1^{\text{Rob}}, \Gamma_2^{\text{Rob}}, \kappa^{\text{Rob}}$ are large enough so that corresponding uncertainty set for this method with these radii contains the true (unknown) CATE. We argue that for any values of $\Gamma_1, \Gamma_2, \kappa$ in Eq. (20), the corresponding value of Γ_2^{Rob} necessary to cover the worst-case realization in Eq. (20) is fairly large. Consequently, this performance bound is quite small, likely negative, and Robust-2 will perform poorly.

To see this, we compute the worst-case realized CATE over Eq. (20) and show that the linear projection of this CATE onto description functions given by the strata necessarily has a large residual. Specifically, we compute the worst-case realization of the CATE over Eq. (20) for the solution given by Robust-2 and a particular choice of Γ_1, Γ_2, κ. We then perform a linear regression of this (candidate) CATE over the description functions of Robust-2 and plot the resulting standard deviation of the residual (which corresponds loosely to Γ_2^{Rob}). Note that the most optimistic case for Robust 2 is given by $\Gamma_2 = \kappa = 0$. For any other values, this worst-case residual can only have larger standard deviation. We plot this standard deviation versus the Adjusted CV in Fig. EC.5. For comparison, we perform the same procedure with Robust Linear and also plot its values. Intuitively, we think this provides a strong intuition for why Robust-2 performs poorly in this

Figure EC.5 Explaining Performance of Robust Methods under the Setting in Section 5.5

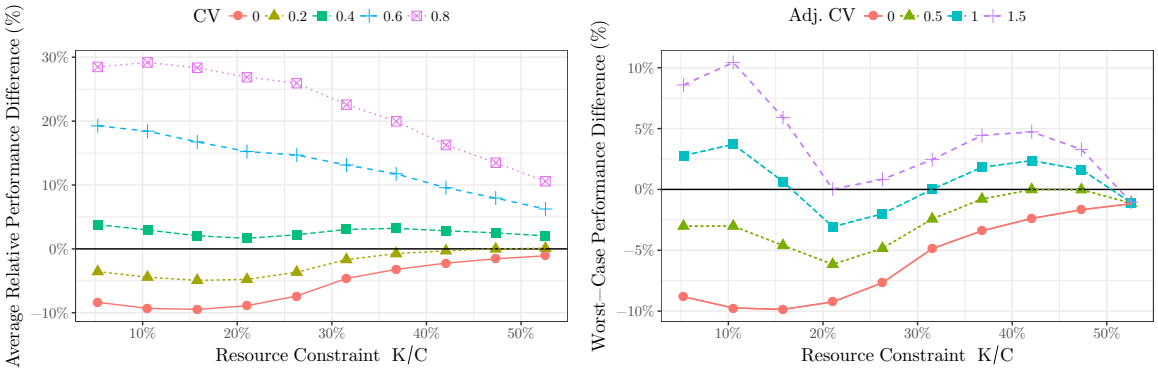


Note. The standard deviation of the above residual is a rough approximation of the value Γ_2^{Rob} required for the uncertainty set of the corresponding robust model to cover the true CATE. Larger values imply poorer performance.

setting; it is highly misspecified, so one needs to accommodate a very large residual to cover the true CATE.

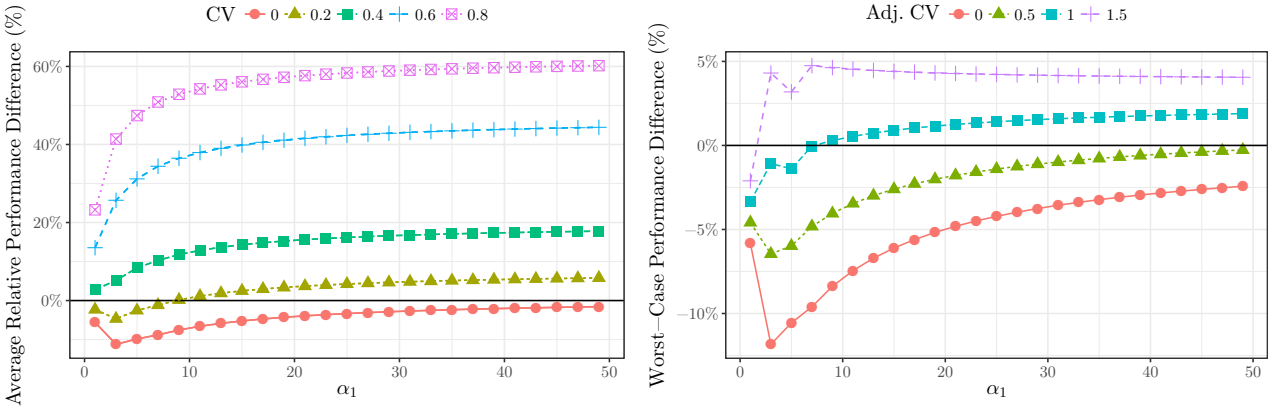
EC.3.5. Graphs from Section 5.6

Figure EC.6 Difference between Robust-2 and Reward Scoring Varying the Resource Constraint



Note. The left panel corresponds to Figure 2 in Section 5.4. The right panel corresponds to Figure 4 Section 5.5.

Figure EC.7 Difference between Robust-2 and Reward Scoring Varying Reward Distribution



Note. The left panel corresponds to Figure 2 in Section 5.4. The right panel corresponds to Figure 4 Section 5.5.

References

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research* 55(1):80–98.

Athey S, Imbens GW (2015) Machine learning methods for estimating heterogeneous causal effects. *Stat* 1050(5).

Athey S, Wager S (2017) Efficient policy learning. *arXiv preprint arXiv:1702.02896* .

Bastani H, Bayati M (2016) Online decision-making with high-dimensional covariates. *SSRN: https://ssrn.com/abstract=2661896* .

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Ben-Tal A, Den Hertog D, Vial JP (2015) Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming* 149(1-2):265–299.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton, New Jersey: Princeton University Press).

Bertsimas D, Copenhaver MS (2017) Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research* .

Bertsimas D, Gupta V, Kallus N (2017) Robust sample average approximation. *Mathematical Programming* 1–66.

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Mathematical Programming* 167(2):235–292.

Bertsimas D, O'Hair A, Relyea S, Silberholz J (2016a) An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science* 62(5):1511–1531.

Bertsimas D, Silberholz J, Trikalinos T (2016b) Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening. *Health Care Management Science* 1–14.

Billings J, Parikh N, Mijanovich T (2000) Emergency department use in New York City: A substitute for primary care? *Issue Brief (Commonwealth Fund)* 433:1–5.

Billings J, Raven MC (2013) Dispelling an urban legend: Frequent emergency department users have substantial burden of disease. *Health Affairs* 32(12):2099–2108.

Bortfeld T, Chan TCY, Trofimov A, Tsitsiklis JN (2008) Robust management of motion uncertainty in intensity-modulated radiation therapy. *Operations Research* 56(6):1461–1473.

Brown DB, Sim M (2009) Satisficing measures for analysis of risky positions. *Management Science* 55(1):71–84.

Chan TCY, Demirtas D, Kwon RH (2016) Optimizing the deployment of public access defibrillators. *Management Science* 62(12):3617–3635.

Chan TCY, Shen ZJM, Siddiq A (2017) Robust defibrillator deployment under cardiac arrest location uncertainty via row-and-column generation. *Operations Research* .

Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases* 40(5):373–383.

Cole SR, Stuart EA (2010) Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology* 172(1):107–115.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612.

Deming D, Dynarski S (2009) Into college, out of poverty? Policies to increase the postsecondary attainment of the poor. Technical report, National Bureau of Economic Research.

Den Hertog D (2018) Is DRO the Only Approach for Optimization Problems with Convex Uncertainty? URL `https://www.birs.ca/events/2018/5-day-workshops/18w5102/videos/watch/201803080905-denHertog.html`.

Deo S, Rajaram K, Rath S, Karmarkar US, Goetz MB (2015) Planning for HIV screening, testing, and care at the Veterans Health Administration. *Operations Research* 63(2):287–304.

Eichler HG, Abadie E, Breckenridge A, Leufkens H, Rasi G (2012) Open clinical trial data for all? A view from regulators. *PLoS Medicine* 9(4):e1001202.

Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. *Medical Care* 36(1):8–27.

Elixhauser A, Steiner C, Palmer L (2014) Clinical Classifications Software (CCS). Agency for Healthcare Research and Quality, URL `https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp`.

Esfahani PM, Kuhn D (2015) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 1–52.

Gao R, Chen X, Kleywegt AJ (2017) Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050* .

Ghaoui LE, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications* 18(4):1035–1064.

Gibbs AL, Su FE (2002) On choosing and bounding probability metrics. *International statistical review* 70(3):419–435.

Goh J, Bayati M, Zenios SA, Singh S, Moore D (2018) Data uncertainty in markov chains: Application to cost-effectiveness analyses of medical innovations. *Operations Research* .

Gutierrez P, Gérardy JY (2017) Causal inference and uplift modelling: A review of the literature. *International Conference on Predictive Applications and APIs*, 1–13.

Hartman E, Grieve R, Ramsahai R, Sekhon JS (2015) From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects. *Journal of Royal Statistical Society: Series A (Statitsics in Society)* 10:1111.

Higgins J, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21(11):1539–1558.

Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–960, ISSN 01621459, URL `http://www.jstor.org/stable/2289064`.

Iancu DA, Trichakis N (2013) Pareto efficiency in robust optimization. *Management Science* 60(1):130–147.

Imai K, Ratkovic M, et al. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1):4–29.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Jackson C, DuBard A (2015) It's all about impactability! Optimizing targeting for care management of complex patients. *Community Care of North Carolina.* Data Brief 4, URL `https://www.communitycarenc.org/media/files/data-brief-no-4-optimizing-targeting-cm.pdf`.

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

Kallus N (2017) Recursive partitioning for personalization using observational data. Precup D, Teh YW, eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1789–1798 (International Convention Centre, Sydney, Australia: PMLR).

Kuhn D, Wiesemann W, Georghiou A (2011) Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming* 130(1):177–209.

Lam H (2016) Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research* 41(4):1248–1275.

Lee KH, Davenport L (2006) Can case management interventions reduce the number of emergency department visits by frequent users? *The Health Care Manager* 25(2):155–159.

Negoescu DM, Bimpikis K, Brandeau ML, Iancu DA (2017) Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science* .

Pardo L (2005) *Statistical inference based on divergence measures* (Chapman and Hall/CRC).

Phillips GA, Brophy DS, Weiland TJ, Chenhall AJ, Dent AW (2006) The effect of multidisciplinary case management on selected outcomes for frequent attenders at an emergency department. *Medical Journal of Australia* 184(12):602.

Shah R, Chen C, O'Rourke S, Lee M, Mohanty SA, Abraham J (2011) Evaluation of care management for the uninsured. *Medical Care* 49(2):166–171.

Shumway M, Boccellari A, O'Brien K, Okin RL (2008) Cost-effectiveness of clinical case management for ED frequent users: Results of a randomized trial. *The American Journal of Emergency Medicine* 26(2):155–164.

Simon HA (1955) A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1):99–118.

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386.

Vizcaíno VM, Aguilar FS, Gutiérrez RF, Martínez MS, López MS, Martínez SS, García EL, Artalejo FR (2008) Assessment of an after-school physical activity program to prevent obesity among 9 to 10-year-old children: A cluster randomized trial. *International Journal of Obesity* 32(1):12.

Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *Journal of Machine Learning Research* 10(Jul):1485–1510.

Zhao Y, Fang X, Simchi-Levi D (2017) A practically competitive and provably consistent algorithm for uplift modeling. *Data Mining (ICDM), 2017 IEEE International Conference on*, 1171–1176 (IEEE).

Zhen J, de Ruiter FJ, den Hertog D (2017a) Robust optimization for models with uncertain SOC and SDP constraints. *Optimization Online PrePrint* URL `http://www.optimization-online.org/DB_HTML/2017/12/6371.html`.

Zhen J, den Hertog D, Sim M (2017b) Adjustable robust optimization via fourier-motzkin elimination. *Operations Research* .

Zuckerman S, Williams AF, Stockley KE (2009) Trends in medicaid physician fees, 2003–2008. *Health Affairs* 28(3):w510–w519.