Characterizing Subpopulations with Better Response to Treatment Using Observational Data - an Epilepsy Case Study

Michal Ozery-Flato, Tal El-Hay, Ranit Aharonov, Naama Parush-Shear-Yashuv, Yaara Goldschmidt

IBM Research Lab - Haifa, Israel

Simon Borghs, Jane Chan, Nassim Haddad, Bosny Pierre-Louis, Linda Kalilani

UCB-Pharma, Brussels, Belgium

Abstract

Electronic health records and health insurance claims, providing observational data on millions of patients, offer great opportunities, and challenges, for population health studies. The objective of this study is to utilize observational data for identifying subpopulations that are likely to benefit from a given treatment, compared to an alternative. We refer to these subpopulations as "better responders", and focus on characterizing them using linear scores with a limited number of variables. Building upon well-established causal inference techniques for analyzing observational data, we propose two algorithms that generate sparse linear scores for identifying better responders, as well as methods for evaluating and comparing such scores. We applied our methodology to a large dataset of ~135,000 epileptic patients derived from claims data. Out of this sample, 85,000 were used to characterize subpopulations with better response to next-generation ("Newer") anti-epileptic drugs (AEDs), compared to an alternative treatment by first-generation ("Older") AEDs. The remaining 50,000 epileptic patients were then used to validate and compare the ability of our scores to identify large subpopulations of epileptic patients with significantly better response to newer AEDs.

1. Introduction

A central problem in population health studies is inferring the influence of a treatment, or intervention, on a given outcome. For example, the influence of the drug "metformin" on the risk for cancer incidence [1]. There are several widely-used statistical techniques for estimating the average effect of a treatment, with respect to a given population [2]. However, the average effect may vary across different sub-populations, due to differences in individual susceptibilities to treatment. Therefore, certain patient groups may show stronger, or alternatively, weaker, response to treatment, than the larger population. In this study we focus on identifying "better responders", which are subpopulations that would benefit more from a given treatment, compared to the larger population. For the remaining population, (i.e. *not* "better responders") the alternative treatment is equally beneficial, or better.

Estimating whether a patient is a better responder to a treatment requires comparing the outcome in two "parallel realities"; one in which the treatment is given, while in the other the alternative is used. We refer to these compared outcomes as "counterfactual", as only one of them can be observed. In other words, for each patient we don't know what would have been the outcome had the alternative been used. Consequently, identifying better responders is essentially different from ordinary prediction problems, since we have no "better response" labels for the patients. Here we focus on finding interpretable models for identifying "better responders", which point to the major *differentiating factors* between better and worse responders. Such models can be used for characterizing subpopulations of better responders. Such models have a large applicability in Health Economics and Outcome Research (HEOR), as well as for building personalized treatment recommendation systems.

An essential element in addressing the task of identifying better responders is the ability to estimate the average causal effect in a given population. Specifically, this enables the validation that a group of patients identified as better responders has indeed a better average causal effect than the larger population. Ideally, average causal effects should be estimated with *interventional studies*, in which the researcher has control over treatment assignment. *Randomized controlled trials*, in which participants are randomly assigned to treatment groups, are the most common design of interventional population health studies. Unfortunately, interventional studies are often costly, sometimes impractical, and may raise ethical questions. On the other hand, real world evidence (RWE) data is abundant and offers new opportunities to study causal effects. Observational studies of RWE data, such as retrospective analysis of electronic

health records and claims data, are challenging alternatives to interventional studies, and are an active research field [3].

The main difficulty in inferring causal relations from observational data is that there is no control over treatment assignment. Furthermore, in many cases, there is no full understanding of the controlling mechanism. There are several common statistical methods for identifying average causal effects in observational data, such as inverse probability weighting and standardization [2,3]. These methods can be used to test specific hypotheses for better responder groups. For example, the hypothesis that a certain anti-depressant drug has a better effect in women [4] may be tested on observational data using such methods. Notably, the common methods for identifying effects are not designed to generate such hypotheses, only to validate them. One may suggest enumerating all hypotheses regarding the characterization of better responders, and test each one using these validation methods. However, when considering multivariate models based on a multitude of variables, this implies a severe multiple testing problem. Therefore, for high dimensional data, such as electronic health records, this approach becomes practically infeasible.

Previous works presented methods for identifying better responders using randomized control trials [5,6]. However, these studies are not directly applicable to observational data due to the inherent biases in the assignment to treatment groups. A different approach that uses observational data to identify patients that may benefit from a treatment is based on identifying patients at high risk for the alternative treatment [7,8]. This approach provides a partial solution to the problem stated above as it focuses on a specific subset and does not address the entire population. It also does not guarantee that these patients will be at a lower risk with the treatment under consideration.

In this work, we combine predictive modeling and causal inference theory to generate scores for identifying better responders using observational data. To ensure a simple and interpretable characterization of the identified subpopulations, we limit the generated scores to be sparse (i.e. including few variables) and linear. The generated scores are validated on a held-out test set, by verifying that the estimated average causal effect for groups of high-scored (respectively, low-scored) patients is larger (respectively, smaller) than the estimated average causal effect in the entire population.

We applied our methodology to a large dataset of epilepsy patients, comparing two alternative classes of anti-epileptic drugs (AEDs): “Newer” vs. “Older”. The first class included second-generation AEDs that were approved over the last two decades for treating epilepsy in the US: felbamate, gabapentin, lacosamide, lamotrigine, levetiracetam, oxcarbazepine, pregabalin, tiagabine, topiramate, vigabatrin, and zonisamide. The second class contained the following first-generation AEDs that are available in the US market: carbamazepine, phenobarbital, phenytoin, primidone, and sodium valproate. In general, AEDs are initially approved as adjunctive therapy for patients with refractory epilepsy, based on data from placebo-controlled trials. Additional indications may be sought to match specific AEDs to patients, but in general there is a dearth of comparative AED “head-to-head” data. When an AED is initially marketed, there is uncertainty regarding the benefit to most epileptic patients, having less severe epilepsy as compared to the available Older AEDs. It is acknowledged within the clinical community that AED selection for epilepsy management should be optimized by adapting the treatment decisions to the characteristics of the individual. The selection should consider both drug factors and patient-specific factors, such as age, sex, childbearing potential, comorbidities, and concomitant medications [9]. Some of these drug and patient-specific factors broadly apply both to Newer and Older AEDs, despite within-category differences. However, not one AED is effective in all cases. Even when applied to a population that theoretically has a high chance to respond to it, prognosis remains difficult in most cases [10]. It is likely that there are characteristics that play an individual or interactive role in determining response / non-response that are currently unknown. This study aims to elucidate some of these characteristics by characterizing better responders to Newer AEDs using retrospective claims data.

2. Material and Methods

We start by formulating the problem of characterizing better response, providing the necessary background and terminology on causal inference concepts (Section 2.1). In Section 2.2 we present two algorithms for this problem. We present methods for evaluating and comparing scores for better response in Section 2.3. Finally, in Section 2.4 we describe the epilepsy use case in which we tested our methodology, and describe specific implementation details. An overview of the entire methodology is presented in Figure 1. All the methods and data analysis described in this paper were implemented using MATLAB 8.1 (R2013a).

2.1. Problem definition

Suppose we have two treatment options, denoted by $a = 1$ and $a = 0$. We assume that only one of these treatment options is given to a patient. Let Y be a random variable that indicates a patient outcome used for evaluating the response to the given treatment. Examples for outcomes may be: death, test result (e.g. blood glucose level), healthcare utilization, etc. We denote by A the random variable that indicates the assigned treatment, and by Y^a the random variable corresponding to the potential outcome when $A = a$. When $A = 1$ then $Y = Y^{a=1}$, and when $A = 0$ then $Y = Y^{a=0}$ is observed. The outcomes $Y^{a=0}$ and $Y^{a=1}$ are referred to as "counterfactuals", as only one of them is observed for each individual. The average causal effect is defined by the deviance between the expected potential outcomes for the two treatment alternatives, $E(Y^{a=1})$ and, $E(Y^{a=0})$. Specifically, when the outcome variable Y is dichotomous (e.g. hospitalized / non-hospitalized, death/survival) the average causal effect is the deviance between the two potential outcome *probabilities*: $p_1 = P(Y^{a=1} = 1)$, and $p_0 = P(Y^{a=0} = 1)$. The deviance between $E(Y^{a=1})$ and $E(Y^{a=0})$ can be measured in several ways, such as taking the difference:

$$E(Y^{a=1}) - E(Y^{a=0}) \tag{1}$$

or the ratio:

$$E(Y^{a=1})/E(Y^{a=0}) \tag{2}$$

For a dichotomous outcome, it is common to consider the *odds ratio* (OR) of $p_1 = P(Y^{a=1} = 1)$ and $p_0 = P(Y^{a=0} = 1)$ as the measure of the effect:

$$\frac{p_1 / (1-p_1)}{p_0 / (1-p_0)} \tag{3}$$

If the assignment of patients to treatments was random (i.e. A is randomly set), then $P(Y^{a=1} = 1)$ could be estimated by $P(Y = 1 \mid A = a)$. Randomized trials use randomization of A for just this purpose. However, in observational data, such as electronic health records or claims data, treatment assignment is usually far from being random. In a real-world setting, treatment assignment (A) often depends on several factors that can potentially affect the outcome (Y). Such factors, which potentially affect both treatment assignment and the outcome, are referred to as *confounders*. We employ the standard *strong ignorability assumption* [11] that potential outcomes are independent of the treatment assignment when conditioned on the covariates X (i.e. no hidden confounders):

$$Y^a \perp A \mid X \quad \text{for } a = 0,1$$

We refer to a set of variables $\boldsymbol{L} \subset \boldsymbol{X}$ as a *sufficient set of confounders* if

$$\boldsymbol{Y^a \perp A \mid L} \quad \text{for } \boldsymbol{a = 0,1}$$

For such set, the expected potential outcome for treatment $a = 0,1$ can be computed by

$$\begin{aligned} E(Y^a) &= \sum_l E(Y^a \mid L = l) * P(L = l) \\ &= \sum_l E(Y^a \mid A = a,\ L = l) * P(L = l) \\ &= \sum_l E(Y \mid A = a,\ L = l) * P(L = l) \end{aligned}$$

The expected potential outcomes, and consequently the average causal effect, may change between different sub-populations, e.g. men vs. women, older vs. younger. We say that a random variable M is an *effect modifier* when the average causal effect varies across different levels of M.

We now formulate the objective of this study, which is to identify better responders using sparse linear scores. We assume the interpretation of the response to be monotonic: a better response corresponds to either higher values or lower values of the response. For simplicity, and without loss of generality, we consider below only the former case (i.e., higher values = better response, while lower values = worse response). Let X be the set of observed variables, excluding the outcome variable Y and the treatment variable A. Given a relatively small number k, the goal is to find a subset $\{M_1, \dots, M_{k'}\} \subseteq X$, $k' \leq k$, and a linear score:

$$f(M_1, \dots, M_{k'}) = \sum_i \alpha_i M_i \tag{4}$$

such that a subpopulation of patients with higher scores will have a better (i.e. larger) value for the average causal effect.

2.2. The algorithms

We propose two different algorithms for generating a linear score for better response with a bounded number of variables. Such scores highlight the major factors that differentiates better responder subgroups from the larger population, facilitating their characterizing. The two algorithms share several common properties: They are both given a sufficient set of confounders, L, and a set of potential effect modifiers, $X' \subseteq X$, as input; Additionally, each of the algorithms uses a stepwise variable selection on X' to generate a sparse linear model for the conditional average effect. The main difference between the two algorithms lies in the way they estimate conditional average effects. The first algorithm learns a prediction model for the outcome, and uses it to estimate the expected causal effect for *each individual*. It then fits a sparse linear regression model to the estimated individual effects. The second algorithm estimates conditional average effects with linear marginal structural models (MSMs) [12]. These linear MSMs yield linear scores for conditional effects. The two algorithms are described in detail below.

.

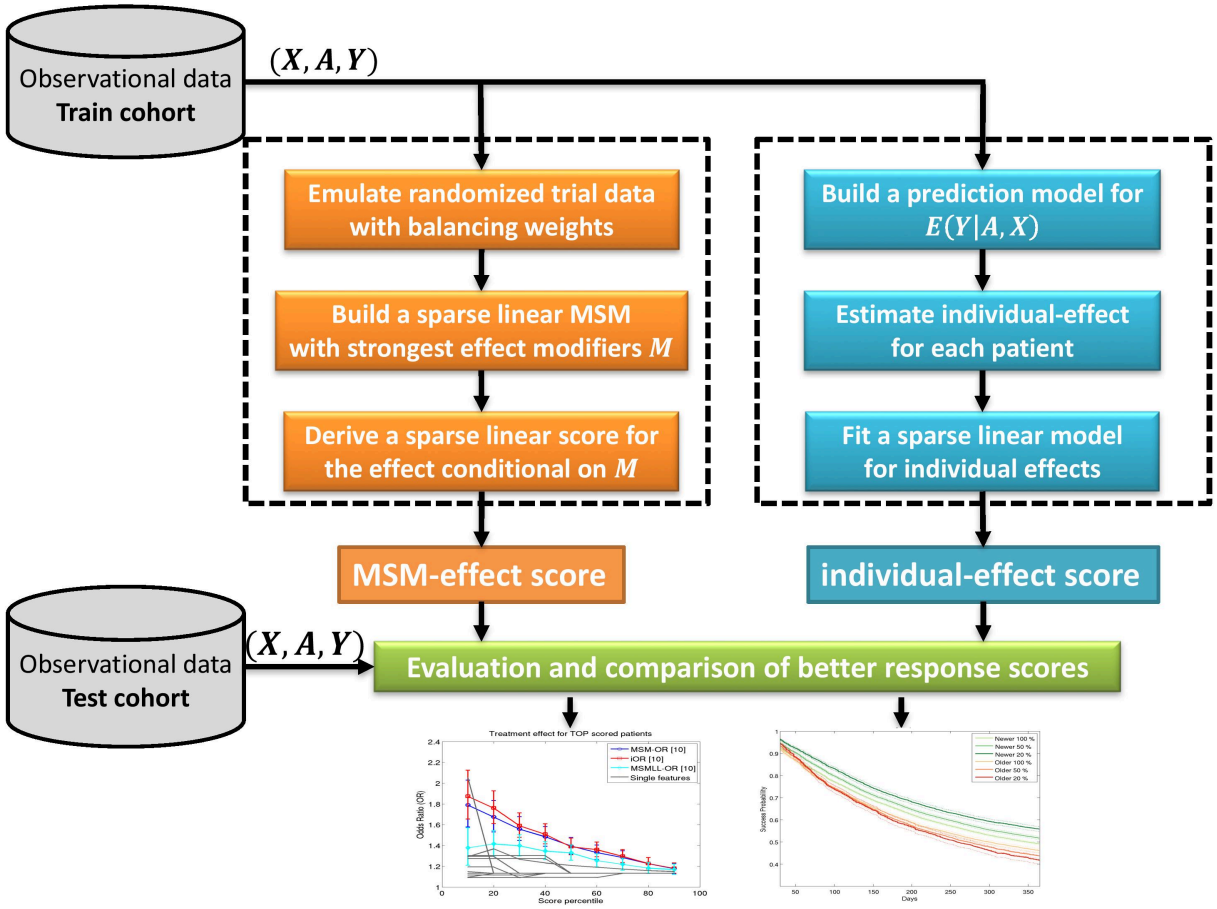

Figure 1: Overview of the methodology for generating and validating sparse linear scores for better responders. The symbols X, A, and Y correspond to observed variables, assigned treatment, and treatment outcome respectively.

The individual-effect score

For a sufficient set of confounders, L, we get $E(Y^{a} \mid L) = E(Y \mid A, L)$. We assume that conditioning on additional variables from X, and in particular on the set of potential effect modifier X', does introduce new biases. Therefore $E(Y^{a} \mid L, X') = E(Y \mid A, L, X')$. Suppose that we have a prediction model for $E(Y \mid A, L, X')$, which generates predictions $\hat{E}(Y \mid A, L, X')$. We can use apply model to each individual in the data and, and use the predicted values $\hat{E}(Y \mid A = 1, L, X')$ and $\hat{E}(Y \mid A = 0, L, X')$ as estimates for $E(Y^{a=1} \mid L, X')$ and $E(Y^{a=0} \mid L, X')$. This allows estimating

the individual effect for each patient in the data, based on his/her own values for L and X'. Finally, we augment patients' data with their estimated individual effects as labels, and fit a sparse linear model that approximates these labels. A formal presentation of this approach is given Algorithm 1 below.

Algorithm 1: individual-effect Score

Input: L - a sufficient set of confounders, X' - potential effect modifiers, k – maximum number of variables

1: Fit a model $\mathcal{M}$ for predicting $E(Y \mid A, L, X')$

2: For each individual $(L = l, X' = x')$:

3: Use the model $\mathcal{M}$ to predict potential outcomes $\hat{E}(Y^{a=1} \mid L, X') = \hat{E}(Y \mid A = 1, L = l, X' = x')$ and $\hat{E}(Y^{a=0} \mid L = l, X' = x') = \hat{E}(Y \mid A = 0, L = l, X' = x')$

4: Use predicted potential outcomes $\hat{E}(Y^{a=1} \mid L = l, X = x')$ and $\hat{E}(Y^{a=0} \mid L = l, X' = x')$ to estimate the individual effect

5: Fit a linear regression model $f(M) = \alpha_0 + \vec{\alpha} \cdot M$, $|M| \leq k$, for predicting the individual effect, using stepwise selection on X'

6: **return** $f(M) = \vec{\alpha} \cdot M$

The MSM-effect Score

Marginal structural models (MSMs) are models for the average potential outcome $E(Y^a)$ [12,3]. MSMs can be extended to include effect modifiers, that is, predict $E(Y^a|M)$ where M corresponds to one or more effect modifiers [12,3]. MSMs use the inverse probability weights (IPW) method to reweight the population, such that in the resulting pseudo-population the treatment assignment variable, A, is independent of the observed variables, X. Additional details on the IPW method are given in Section 2.4. MSMs are fitted in a two-stage process. In the first stage, individuals' weights are computed, such that there are no confounders in the reweighted population. In the second stage, an outcome prediction model is fit to the reweighted population.

Let $M \subseteq X'$ be a set of effect modifiers. If the causal effect is measured by the difference, $E(Y^{a=1}) - E(Y^{a=0})$, then we use the following linear MSM:

$$E(Y^a \mid M) = \psi_0 + \psi_1 a + \vec{\psi}_2 \cdot Ma + \vec{\psi}_3 \cdot M \tag{5}$$

where $\vec{\psi}_2$ and $\vec{\psi}_3$ are coefficient vectors at the size of M. From this MSM we obtain a linear model for the conditional effect:

$$E(Y^{a=1} \mid M) - E(Y^{a=0} \mid M) = \psi_1 + \vec{\psi}_2 \cdot M \tag{6}$$

If the ratio $E(Y^{a=1})/E(Y^{a=0})$ is used for measuring the effect, then we consider a linear MSM with log link function:

$$\log E(Y^a \mid M) = \theta_0 + \theta_1 a + \vec{\theta}_2 \cdot Ma + \vec{\theta}_3 \cdot M \tag{7}$$

This MSM leads to a linear model for the *log* of the conditional effect

$$\log E(Y^{a=1}|M)/E(Y^{a=0}|M) = \log E(Y^{a=1} \mid M) - \log E(Y^{a=0} \mid M) = \theta_1 + \vec{\theta}_2 \cdot M \tag{8}$$

Finally, when using the odds-ratio for measuring the effect for a dichotomous outcome Y, we use the following linear MSM with logit link function

$$\text{logit } P(Y^a = 1 \mid M) = \beta_0 + \beta_1 a + \vec{\beta}_2 \cdot Ma + \vec{\beta}_3 \cdot M \tag{9}$$

The odds-ratio measurement of the effect conditioned on M is:

$$OR(M) = \frac{p_{1,M} / (1-p_{1,M})}{p_{0,M} / (1-p_{0,M})}$$

Where $p_{a,M} = P(Y^a = 1 \mid M)$. We use the MSM in Equation 9 to obtain a linear model for the *log* of the conditional effect

$$\begin{aligned} \log OR(M) &= \log p_{1,M}(1 - p_{1,M}) - \log p_{0,M}(1 - p_{0,M}) \\ &= \text{logit}\, P(Y^{a=1} = 1 \mid M) - \text{logit}\, P(Y^{a=0} = 1 \mid M) \\ &= \beta_1 + \vec{\beta}_2 \cdot M \end{aligned} \tag{10}$$

For each of the three causal effect measures that we consider: difference, ratio and odds-ratio, the corresponding linear MSM yields a linear score that estimates the conditional effect, or the log of it. Either way, the resulting score preserves the ranking of the patients induced by the conditional effect estimations predicted by the MSM. Consequently, higher scores correspond to subpopulations with larger estimated effect values.

A variable is said to have *additive effect modification* if the corresponding coefficient in $\vec{\psi}_2$ / $\vec{\theta}_2$ / $\vec{\beta}_2$ is significantly different than 0. The number of variables that we include in the MSM is limited, and hence we would like to select those having maximal additive effect modification. The MSM-effect score uses the greedy heuristic of stepwise variable selection, adding in each iteration a *pair* of terms $(x' + x'a)$ to the MSM, where $x' \in X'$ has a maximal additive effect modification, if such exists. See Algorithm 2 below for complete details on the MSM-effect score.

Algorithm 2: MSM-effect Score

Input: L - a sufficient set of confounders, X' - potential effect modifiers, k – maximum number of variables

Use L to compute weights such that the reweighted population has no confounders

$M \leftarrow \emptyset$

Iterate k times:

    For each variable $X_i \in X'$:

        Fit a linear MSM model with the set of variables $M \cup \{X_i\}$
        (see Equations 5, 7, 9 for difference, ratio and odd-ratio measures of causal effect)

        Evaluate the additive effect modification of X_i in this MSM model using the p-value of the coefficient corresponding to the product term $X_i * A$ to be different from 0.

    If at least one of the variables has an additive effect modification significantly different from 0:

        $M \leftarrow M \cup \{X_{imax}\}$ where X_{imax} is a variable having the most significant additive effect

    Else: stop the iteration

return $f(M) = \vec{\alpha} \cdot M$, where $\vec{\alpha}$ is the coefficient vector corresponding to the product terms in the final MSM. ($\vec{\alpha} = \vec{\psi}_2$ in Equation 5, $\vec{\alpha} = \vec{\theta}_2$ in Equation 7, $\vec{\alpha} = \vec{\beta}_2$ in Equation 9.)

2.3. Scores Evaluation and Comparison

The evaluation and comparison of the generated scores is done on a held-out test set. A score $f(M)$ is expected to be an effect modifier since the average causal effect should vary across different levels this score. We verify that a score $f(M)$ is an effect modifier by testing whether the corresponding random variable has an additive effect modification in the MSM $E(Y^a \mid f(M))$. To compare scores, we plot curves that map every score-percentile to the average causal effect computed within the corresponding group of individuals (i.e. top or bottom-scored individuals defined by that percentile). Ideally, higher score percentiles should correspond to larger average causal effect. For dichotomous outcomes that correspond to events, such as hospitalization and death, we analyze the corresponding time for these events. We use Kaplan-Meier curves to compare the distribution of time-to-event for the two potential outcomes (i.e. $a = 1,0$) corresponding to the treatment and its alternative. To account for the bias in treatment assignment, the Kaplan-Meier curves are computed on a reweighted population in which there is no bias between the two treatment

groups (i.e. no confounders). We repeat this comparison for different scores percentiles, to verify that high-scored (respectively, low-scored) patients have larger (respectively, smaller) values for time-to-event.

We consider a variable selected to a score $f(M)$ as robust, if it is likely to be selected on a different sample of the data. We test the robustness of the variables selected for each score using non-parametric bootstrapping. This is done by sampling the data with replacement, generating a new dataset of the same size as the original dataset (i.e. the same number of patients). We perform N iterations of bootstrapping, generating our scores for each dataset. The robustness of a selected variable is evaluated by the fraction of the times it is selected during N bootstrapping iterations.

2.4. The epilepsy case study

The Data

To test our methodology, we used a dataset of ~135,000 epileptic patients derived from the IMS Health Surveillance Data Incorporated (SDI) medical claims database. The SDI database contains medical, pharmacy, and hospital claims collected from all major regions of the United States between January 1, 2006 and September 31, 2011. Epileptic patients were identified based on their diagnoses and prescribed drugs. For a description of the inclusion criteria we used for epileptic patients see Appendix A.

Every patient in our dataset was assigned with an index-date, which is the start of an AED treatment having exactly one added drug that was not in the previous AED regimen. We refer to this added drug as the *index drug*. The year before the index-date is referred as the *baseline period*. It was used to derive variables (X) that characterize the patient in that time period. The year after the index-date is referred as the *treatment evaluation period* and was used to compute the outcome of the treatment (Y). See Figure 2 for a schematic explanation of the referred time periods and their relation to variable computations.

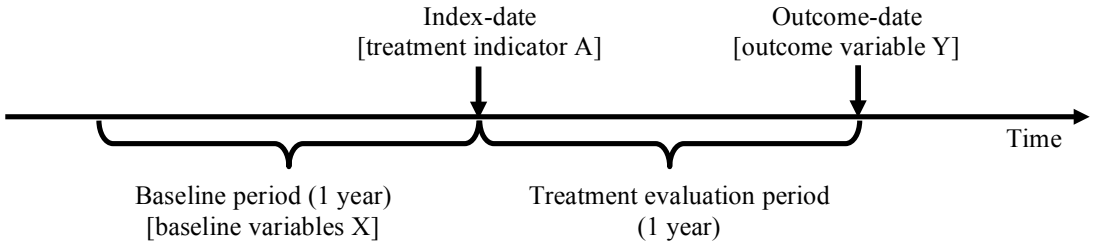

Figure 2: Time period definition. Square brackets contain variables that are computed at that time point/period.

The treatment starting at the index-date was classified as "Newer AED" ($A = 1$) or "Older AED" ($A = 0$) based on the class of the index drug. The evaluation of a treatment was done in the year following the index-date. Because the primary symptom of epilepsy - seizures – was not available in the SDI database, we used treatment changes as a proxy measure of seizure control and patients status. An unsuccessful outcome ($Y = 0$) was defined as any change other than a dose change (i.e., increase/decrease) or a complete withdrawal of any AED treatment in the subsequent 1 to12 months after the index-date. A longer-term stable treatment or a complete withdrawal from an AED therapy was considered a successful outcome ($Y = 1$).

The derived variables (X) included the following: age (at index date); gender; type of treatment change at index date; epileptic-state variables (indicator for each epilepsy-related ICD9 code, counts for each epilepsy related CPT code, proxies for generalized/focal epilepsy as well as proxies for seizures[13]); consumption of AEDs (mean possession ratio (MPR) and indicator for any use of AED and of each AED, indicator if patient used Newer/Older AEDs, the number of different AEDs, the number of treatment change events); comorbidities (based on an adapted list [14], as well as by diagnostic codes); mean monthly activity (using diagnoses, prescriptions and hospitalization data), indicator for hospital encounters, non-AED treatments (indicators for specific list of non-epilepsy drugs[14]), ecosystem variables (payer, state, first digit of zip code, specialty of physicians, year of index date).

Finally, the dataset was randomly partitioned into two sets: train and test datasets, which totaled ~85,000 and 50,000 patients, respectively.

The effect measure and scoring algorithms

Since we had a dichotomous outcome variable, we used the odds-ratio (OR) as a measure of the effect. We refer to the OR variant of the MSM-effect and individual-effect algorithms as MSM-OR and individual-OR respectively. We use the term iOR as an abbreviation for individual-OR. We generated the MSM-OR and iOR scores using the train dataset, and evaluated and compared these scores on the test dataset.

Potential Confounders and Effect Modifiers

A confounder variable is a shared cause of both the outcome Y and the assigned treatment A, therefore it may have a statistical correlation with each. Identifying potential confounders is a key problem in causal analysis of observational studies [15]. In the epilepsy case study, we identified a set of potential confounders, L, using the following ad-hoc procedure. We first excluded nearly constant variables (mode frequency larger than 0.99). In accordance with a previous recommendation [16] and to avoid overfitting of our models for $P(A \mid L)$ and $P(Y \mid A, L)$, we included in L only variables that were significantly ($P < 0.05$) associated with Y. The association was measured by Chi-square (dichotomous variables) and t-test (continuous variables) after Bonferroni correction for multiple testing. Note that a variable is individually selected based on the strength of its association with Y, that is, without any reference to the association p-values computed for the other variables. Finally, we filtered out from L variables that are highly correlated (Pearson correlation > 0.99) with each other, as such variables are expected to have low additive predictive value. We considered the resulting set of variables in L as a sufficient set of confounders. Since effect modifiers are also expected to be statistically associated with the outcome, Y, we used L as the set of potential effect modifiers and limited the stepwise selection procedure of the two algorithms, MSM-OR and iOR, to select variables only from $X' = L$.

Outcome prediction model for iOR

For the outcome prediction model in the iOR algorithm (see step 1 in Algorithm 1) we used the following logistic regression model:

$$\text{logit } P(Y = 1 \mid A, L, X') = \alpha_0 + \sum_{x_i \in L \cup X'} \alpha_i x_i + \beta_0 A + \sum_{x_i \in X'} \beta_i A x_i \quad (11)$$

In general, other classifiers could be used as well for predicting the outcome, including random forest, SVM, adaBoost etc. Specifically, it is possible to make the prediction models more flexible by considering adding polynomial terms and additional interaction terms. Note, however, that it is unclear how to select the "best" outcome model, since we have missing labels for half of the potential outcome (i.e. ones corresponding to opposite treatment assignment).

Generating balancing weights

In the epilepsy case study, we used the IPW method [2,3] to generate balancing weights for the MSM-OR score (step 1 in Algorithm 2), as well as for evaluating and comparing the scores (Section 2.3). The IPW method creates a pseudo-population from the original one by assigning for every individual with observed treatment value $A = a$ and observed confounder values $L = l$, the following weight:

$$\frac{P(A=a)}{P(A=a \mid L=l)} \quad (12)$$

This re-weighting of the individuals in the original population balances the biases in the distribution of L between the two treatment groups, mimicking randomization and creating a synthetic sample in which the distribution of observed variables is independent of the treatment assignment. The overall sum of weights in the pseudo-population is equal to the number of patients in the original one. The value $P(A = a \mid L = l)$ for each individual was estimated using a simple logistic regression model that was fitted to the data. Similar to the outcome prediction model, one could have considered other types of models for evaluating $P(A = a \mid L = l)$. Models for $P(A = a \mid L = l)$ used in IPW should be selected based on their ability to minimize the bias between treatment groups, and not based on their accuracy (e.g. c-statistic/area under ROC curve) [2]. In the next section we describe a standard statistical method for testing univariate imbalances between treatment groups. In the recent years, additional methods were proposed to improve the balancing by the generated weights [17–22], most of which fit balancing weights directly without modeling $P(A = a \mid L = l)$. Such methods can be incorporated into our framework, instead of the IPW method.

Testing imbalances between treatment groups

After generating a pseudo-population using IPW, (e.g., to compute the OR when working with MSMs) we validated that all observed variables show no imbalances between the two treatment groups. We quantified the balance for each

variable using its standardized difference, d, which is the (absolute) difference in the variable means between the two treatment groups, divided by the combined standard deviation. To be exact, we used the following definition of the standardized difference,

$$d = \frac{|\mu_1 - \mu_0|}{\sqrt{(s_1^2 + s_0^2)/2}} \quad (13)$$

where μ_1 and μ_0 denote the mean in the two treatment groups; s_1 and s_0 denote the sample variance of the variable in two treatment groups. We considered a variable as balanced if its standardized difference was below 0.1.

3. Results

In this section, we present the results of applying our methodology to the dataset of epileptic patients. We start with providing some descriptive statistics on the dataset.

3.1. Data Statistics

Some selected statistics of the train and test datasets are given in Table 1, demonstrating that the two datasets share the same data distributions. To correct for the biases in treatment assignment, we used IPW to create two pseudo populations corresponding to the train and test datasets. In both pseudo-populations, all observed variables were shown to be balanced between the two treatment groups. The single exception was 'index-date year' in the test pseudo-population, which showed a minor imbalance (d=0.12). This variable was not found to be significantly associated with the outcome variable Y in the test dataset and hence was not detected as a possible confounder. The corrected OR values, which were independently computed in the train and test pseudo-populations, indicated that Newer AEDs had a positive casual effect on the outcome. The uncorrected OR values, which were computed in the original train and test datasets, erroneously indicated no causal effect. This striking difference between the correct and uncorrected OR values exemplifies the importance of correcting for the biases in treatment assignment A.

Table 1: Train and test data statistics

Characteristic	Train dataset	Test dataset
Size	83184	50000
Index-drug is Newer (A=1)	74 %	74 %
Successful treatments (Y=1)	49 %	49 %
Gender, female	62 %	62 %
Age at index date (years)	51+-16	51+-16
Index-drug prescribed by neurologist	42 %	42 %
Index-treatment is monotherapy	57 %	57 %
Odds ratio (OR) (corrected by IPW)	1.10 [1.07-1.14], p=1e-9	1.14 [1.09-1.18], p=4e-10
Odds ratio (OR) – without correction	0.98 [0.95-1.01], p>0.1	0.99 [0.95-1.03], p>0.1

The total number of variables in X was 682. We selected the set of potential confounders by applying the methodology in Section 2.3. We first excluded 308 nearly constant variables. Of the 374 remaining variables, 173 variables were found to be significantly associated with the outcome Y after Bonferroni correction. Finally, we excluded five additional variables due to high correlation with other variables. Overall, the set of potential confounders L included 168 variables. As described above, L was also used as the set of potential effect modifiers, X'.

3.2. The Scores

We generated the MSM-OR and iOR scores for $K = 10$ on the train dataset. Table 2 presents the variables selected by each of the scores, as well as their coefficients and the number of times they were selected in 10 bootstrap runs. The sets of variables selected by the 2 scores largely overlap, sharing 9/10 of the variables. There were 5 variables

that were selected in 50% or more of the bootstrap runs, by either MSM-OR and iOR algorithms. All these "robust" variables were selected by both MSM-OR and iOR when running on the original train dataset (see Table 2).

Unless stated differently, all the variables in Table 2 were computed for the entire baseline period (one year). For example, the variable "Had neurological comorbidity dx" indicates whether the patient had at least one diagnosis of neurological comorbidity during the year before the index-date.

Table 2: The variables selected for the MSM-OR and iOR scores. The first number in each column indicates the coefficient, while the second number (in brackets) contains the number of times the variable was selected in 10 bootstrap trials.

Selected variables	MSM-OR	iOR
Index-date treatment switched a previous drug	0.3 (8 / 10)	0.4 (10 / 10)
Had neurological comorbidity dx	0.3 (4 / 10)	0.3 (7 / 10)
Was treated by a neurologist	0.2 (6 / 10)	0.2 (4 / 10)
Pregabalin medication MPR**	-- (5 / 10)	0.1 (1 / 10)
Older AED MPR**	0.1 (0 / 10)	-- (0 / 10)
Age	-- (1 / 10)	-0.1 (5 / 10)
Had a seizure proxy* in the previous month	0.6 (10 / 10)	0.7 (10 / 10)
Received Older AED	-0.3 (5 / 10)	-0.2 (2 / 10)
Had Medicare	-0.1 (2 / 10)	-- (4 / 10)
Had disorders of lipid metabolisms dx	-0.2 (4 / 10)	-0.2 (3 / 10)
Had back problem dx	-0.2 (6 / 10)	-0.2 (2 / 10)
Had trauma-related dx	-0.5 (0 / 10)	-0.4 (1 / 10)

* A seizure is inferred by a claim for ER, hospitalization or ambulance with epileptic/seizure primary or secondary diagnosis.
** MPR = medication possession ratio

The variable "Had a seizure proxy in the previous month" was the strongest variable in the two scores. It had the largest coefficient, and was shown to be most robust since it was selected by the two scores in all bootstrap runs. Another variable that was selected by the two scores and was found to be very robust was: "Index-drug switched a previous drug".

3.3. Scores Evaluation

In this section, we evaluate and compare between the iOR and MSM-OR scores. We tested the additive effect modification of the variables corresponding to the two scores, as described in Section 2.3. We showed that both score variables had a significant additive effect modification, with P-values of 7e-36 and 5e-36 for the iOR and MSM-OR, respectively.

We also tested a more intuitive variant of the MSM-OR, which differs from MSM-OR in its optimization objective: Instead of selecting a variable with the most significant additive effect modification, it selected a variable that maximized the overall likelihood of $P(Y \mid A, M)$. We also showed that this variant of the MSM-OR score, which we noted MSMLL-OR, had a significant additive effect modification, although lower than the other two scores, with a P-value of 7e-17.

Figure 3 compares the iOR and MSM-OR scores by plotting the average causal effect, measured by the OR, in top- and bottom-scored patient groups as a function of the score percentile used to identify these groups. As shown in Figure 3, the iOR and MSM-OR scores had very similar performance. They both managed to identify large subpopulations of patients that have significantly larger, or smaller, OR values compared to the OR observed in the entire population. On the other hand, the MSMLL-OR was less successful in identifying groups with significant higher, or lower, OR.

We tested the variables in Table 2 as single-variable scores, and compared them to the previous multi-variable scores. As can be seen in Figure 3, the gray lines that correspond to these variables were not able to significantly identify better treatment-responders – with the expectation of the variable "Had a seizure in the previous month". Since this variable is dichotomous, it was able to identify one group of better treatment responders, which included 11% of the patients.

We compared the time-to-failure in different score-groups, where a failure corresponded to a treatment change. Figure 4 presents IPW-corrected Kaplan-Meier curves for the top-20%, top-50%, and top-100% (i.e., entire population) score-groups for the iOR score. As expected, the time-to-failure was longer on average for Newer AEDs in these score groups; this difference increased for score-groups with higher scores. We repeated the same analysis for bottom-score groups as well as for the MSM-OR score. In accordance to Figure 3, the difference between Older and Newer in bottom-scored groups was in the expected direction (those on Newer AEDs having shorter time-to-failure than those on Older AEDs) but was less pronounced than the difference between top-scores patients (results not shown).

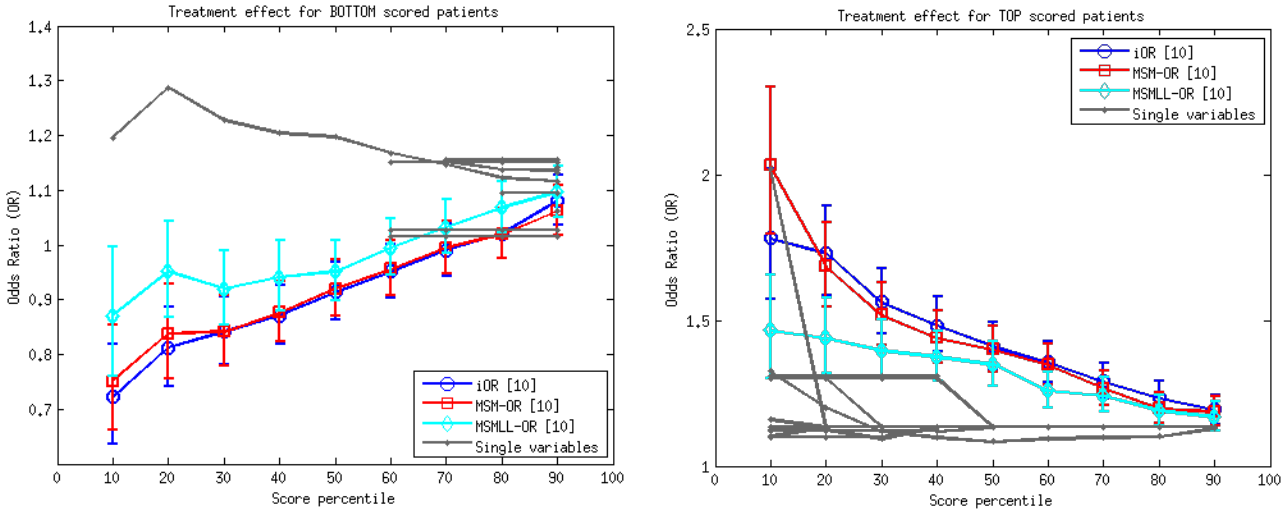

Figure 3: Scores evaluation and comparison. Gray lines refer to single-variable scores corresponding to the 11 variable that were selected to the MSM-OR and/or iOR scores.

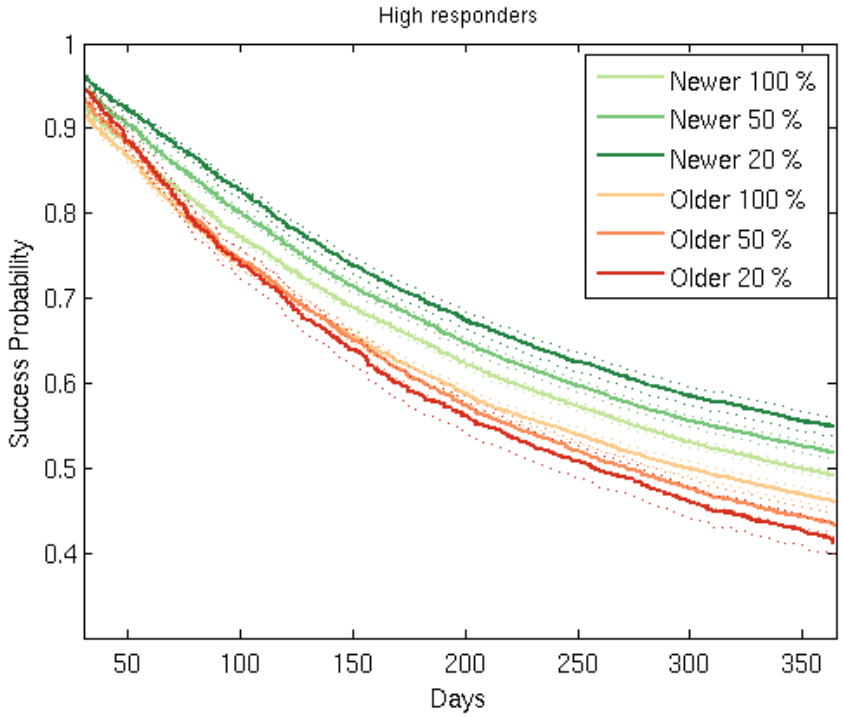

Figure 4: A comparison of Kaplan-Meier curves of time-to-failure for Newer and Older AED in top-scored patients by the iOR score. Newer/Older 20% means those on Newer/Older AEDs from the top-20% score groups and analogously for the 50% and 100% lines. Dotted lines represent 95% confidence intervals.

4. Discussion

This study uses real world observational health data to construct linear scores for identifying patients that are likely to have better response to a given treatment, compared to an alternative, using only a small subset of the available variables. This task, which involves estimating the deviance between the potential outcomes for two alternative treatments, is fundamentally different from the task of identifying subpopulations with better outcomes. The later task can be addressed with standard predictive modeling tools, whereas identifying better responders requires the use of causal inference methodologies to adjust for the inherent biases in the data. The bound on the number of variables in the modeling of better response subgroups requires the integration of sparse learning techniques with causal modeling.

Building on well-established concepts and methods from causal inference and machine learning, we presented two algorithms for identifying subpopulations with differential response to treatments using observational data: individual-effect and MSM-effect. The two algorithms are designed to select the major factors that predict the differential response for the given treatments. The differential response by itself is not given, and therefore traditional variable selection techniques, which require labeled data, cannot be directly applied. The two algorithms take different approaches for this hurdle. The individual-effect algorithm augments the train dataset with estimates of the individual effect for each patient, thus reducing the problem into an ordinary prediction task. The MSM-effect takes a different approach, utilizing the causal inference method of linear MSMs to derive linear scores with strongest effect modifiers.

We applied the odds-ratio variants of the two algorithms, iOR and MSM-OR, to a dataset of epilepsy patients to compare two alternative anti-epileptic treatments: “New” vs. “Old”. To evaluate the generated scores, we employed a machine learning train/test paradigm. The two algorithms yielded similar results both in terms of the set of the selected variables (Table 2), as well as in the ability to identify subgroups of better /worse responders in a held-out dataset (Figure 3). Due to the inherent difference in their modeling assumptions: the iOR algorithm modeled $P(Y \mid A, L)$ while the MSM-OR algorithm modeled $P(A \mid L)$; these algorithms are *not* a-priori guaranteed to produce similar results. If there was a major difference in the results of the two scoring algorithms, it would indicate that at least one of our models was misspecified. Thus, the agreement between the algorithms’ results strengthens their validity and the confidence in the underlying models.

Another difference between the iOR/individual-effect and the MSM-OR/MSM-effect algorithms lies in the intricacy of the variable selection procedure. In this aspect, the individual-effect score has a clear advantage. The individual-effect score uses a standard stepwise procedure that in each step adds a *single* term to a *linear regression* model. This procedure could have been replaced by other variable selection methods, such as Lasso [23]. Conversely, in each step of its stepwise selection procedure the MSM-effect score considers adding a *pair* of terms to a *generalized linear regression* (GLM) model. The running time of the two algorithms is dominated by these stepwise selection procedures, which involve fitting multiple different models in each step. Fitting a linear regression model has a simple closed-form solution, which can be implemented in $O(nd^2)$ operations, where n is the number of samples and d is the number of variables [24]. In contrast, fitting a log regression model or a logistic regression model involves an iterative method that maximizes the likelihood function [25]. Fitting generalized linear models in Matlab (`glmfit` function) and in R (`glm` function) is implemented with the iteratively reweighted least squares (IRLS) method, which takes $O(nd^2)$ operations per iteration [26]. The number of iterations for fitting each model depends on the convergence rate for the data. Overall, in our epilepsy case study, the MSM-OR score was three-time slower than the iOR score (results not shown).

In the epilepsy case study, the iOR and MSM-OR scores were trained to include 10 variables. In comparison to single-variable scores, these two scores were much more successful in identifying better and worse responders (Figure 3). The binary variable "Had a seizure proxy in the previous month" identified a single strong group of better responders totaling ~10% of the patients. Conversely, our iOR and MSM-OR scores could identify much larger groups of better responders in various sizes, with up to ~50% of the patients. Another interesting point is the ability of our scores to identify *worse* responders, that is, patients that are more likely to benefit from Older AEDs. While Older AEDs had a negative effect in the entire dataset, our scores identified a group with ~10% of the patients for which the odds for success were significantly lower for Newer AEDs than for Older AEDs' (OR = 0.77 [0.68-0.87], P-value=5e-05). We note that the significance of the effect for this identified group of worse responders is much less pronounced than the effect observed in groups identified as better-treatment responders. For example, the treatment effect measured in the group corresponding to the 10% top-iOR scores was OR=1.87 [1.65-2.12], P-value= 1e-22.

Most of the variables in Table 2 relate to epilepsy and usage of AEDs. The occurrence of seizures and the existence of comorbidities are known to affect AEDs response. Variables that have been less clearly described in the past, include age and the related variable 'Had Medicare'. Since AEDs are also prescribed for pain relief, it is possible that the selection of the variables 'Had back problem dx', and 'Had trauma-related dx' is a result of a contamination of our dataset with patients that consume AEDs for pain management. This is a drawback of using claims data, which do not make an explicit link between prescriptions and the diagnoses/medical conditions for which they were subscribed.

There are several limitations to our study. In general, observational data studies are limited by the possible existence of unobserved confounders and selection bias. The claims data we analyzed were missing important data relevant to our study, such as seizure frequency, etiology, genetic data and/or neurological test results. Selection bias may exist due to the choice of patients showing regular medical activity. This selection of patients was done to control for potential data gaps due to the open nature of the database. Another limitation in our methodology relates to the selection of confounders, which were selected by their statistical association with the outcome, in accordance with previous recommendations [16]. It may be preferable to base the detection of confounders on domain knowledge and causal diagrams that describe cause and effect relationships between variables [27]. However, identifying such a set in the presence of high-dimensional data where domain knowledge cannot capture the complex structure of the system is a challenging task of practical importance.

This study focused on modeling the differential response with sparse linear scores, which are commonly used in the medical domain for risk prediction. The use of sparse linear models facilitates the understanding of the major "risk factors" for a differential response and their contribution to it. As demonstrated in the epilepsy case study, the set of major predictors for the differential response, as computed by iOR and MSM-OR scores, may be different from the set of major predictors for the (observed) outcome itself, as computed by the MSMLL-OR algorithm. This difference is manifested in the poorer performance of MSMLL-OR in identifying better responders, compared to the iOR and MSM-OR (see Figure 3). The sparsity condition, which limits the number of variables used, may be applied to other models than linear scores. A future research direction is to consider additional (interpretable) models for the differential response, such as linear scores with interactions and decision trees. Such models can be evaluated and compared using the train/test framework we described in Section 2.3. Another direction for future work is to improve the intermediate models of $E(Y \mid A, L)$, in the individual-effect algorithm, and $P(A \mid L)$ in the MSM-effect algorithm. Here we used a simple logistic regression for both models. In principle, other prediction models, such as Random Forest [28], may be considered for these intermediate learning tasks. Note that selecting the best model for a causal

inference task is a challenge by itself, as the ground truth is unknown. A common approach to address this challenge is to test and compare the different models using simulated data under various mechanisms for generating the potential outcomes [29]. An interesting future work is to adapt such simulations for testing and comparing sparse models for differential response.

Acknowledgements

We thank Chen Yanover and two anonymous referees for their helpful comments on the manuscript.

References

[1] J.M.M. Evans, L.A. Donnelly, A.M. Emslie-Smith, D.R. Alessi, A.D. Morris, Metformin and reduced risk of cancer in diabetic patients, BMJ. 330 (2005) 1304–1305. doi:10.1136/bmj.38415.708634.F7.

[2] P.C. Austin, An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies, Multivariate Behav Res. 46 (2011) 399–424. doi:10.1080/00273171.2011.568786.

[3] Hernán MA, Robins JM (2016). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming., n.d.

[4] M. Vermeiden, W.W. van den Broek, P.G.H. Mulder, T.K. Birkenhäger, Influence of gender and menopausal status on antidepressant treatment response in depressed inpatients, J. Psychopharmacol. (Oxford). 24 (2010) 497–502. doi:10.1177/0269881109105137.

[5] T. Cai, L. Tian, P.H. Wong, L.J. Wei, Analysis of randomized comparative clinical trial data for personalized treatment selections, Biostatistics. 12 (2011) 270–282. doi:10.1093/biostatistics/kxq060.

[6] L. Zhao, L. Tian, T. Cai, B. Claggett, L.J. Wei, Effectively Selecting a Target Population for a Future Comparative Study, J Am Stat Assoc. 108 (2013) 527–539. doi:10.1080/01621459.2013.770705.

[7] H. Janes, M.S. Pepe, Y. Huang, A framework for evaluating markers used to select patient treatment, Med Decis Making. 34 (2014) 159–167. doi:10.1177/0272989X13493147.

[8] M.H. Erder, J.E. Signorovitch, J. Setyawan, H. Yang, K. Parikh, K.A. Betts, J. Xie, P. Hodgkins, E.Q. Wu, Identifying patient subgroups who benefit most from a treatment: using administrative claims data to uncover treatment heterogeneity, J Med Econ. 15 (2012) 1078–1087. doi:10.3111/13696998.2012.689270.

[9] E. Perucca, T. Tomson, The pharmacological treatment of epilepsy in adults, Lancet Neurol. 10 (2011) 446–456. doi:10.1016/S1474-4422(11)70047-3.

[10] I. Gilioli, A. Vignoli, E. Visani, M. Casazza, L. Canafoglia, V. Chiesa, E. Gardella, F. La Briola, F. Panzica, G. Avanzini, M.P. Canevini, S. Franceschetti, S. Binelli, Focal epilepsies in adult patients attending two epilepsy centers: Classification of drug-resistance, assessment of risk factors, and usefulness of "new" antiepileptic drugs, Epilepsia. 53 (2012) 733–740. doi:10.1111/j.1528-1167.2012.03416.x.

[11] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika. 70 (1983) 41–55. doi:10.1093/biomet/70.1.41.

[12] J.M. Robins, M.Á. Hernán, B. Brumback, Marginal Structural Models and Causal Inference in Epidemiology, Epidemiology. 11 (2000) 550–560.

[13] N. Shcherbakova, K. Rascati, C. Brown, K. Lawson, S. Novak, K.M. Richards, L. Yoder, Factors associated with seizure recurrence in epilepsy patients treated with antiepileptic monotherapy: A retrospective observational cohort study using US administrative insurance claims, CNS Drugs. 28 (2014) 1047–1058. doi:10.1007/s40263-014-0191-1.

[14] B. Legros, P. Boon, B. Ceulemans, T. Coppens, K. Geens, H. Hauman, L. Lagae, A. Meurs, L. Mol, M. Ossemann, K. van Rijckevorsel, M. Van Zandijcke, P. Vrielynck, D. Wagemans, T. Grisar, Development of an electronic decision tool to support appropriate treatment choice in adult patients with epilepsy--Epi-Scope(®), Seizure. 21 (2012) 32–39. doi:10.1016/j.seizure.2011.09.007.

[15] M. Brookhart, T. Stürmer, R. Glynn, J. Rassen, S. Schneeweiss, Confounding control in healthcare database research: challenges and potential approaches, Med Care. 48 (2010) S114–S120. doi:10.1097/MLR.0b013e3181dbebe3.

[16] M.A. Brookhart, S. Schneeweiss, K.J. Rothman, R.J. Glynn, J. Avorn, T. Stürmer, Variable selection for propensity score models, Am. J. Epidemiol. 163 (2006) 1149–1156. doi:10.1093/aje/kwj149.

[17] J. Hainmueller, Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies, Political Analysis. 20 (2012) 25–46. doi:10.1093/pan/mpr025.

[18] B.S. Graham, D.X. Pinto, C. Campos, D. Egel, Inverse Probability Tilting for Moment Condition Models with Missing Data, Rev Econ Stud. 79 (2012) 1053–1079. doi:10.1093/restud/rdr047.

[19] K. Imai, M. Ratkovic, Covariate balancing propensity score, J. R. Stat. Soc. B. 76 (2014) 243–263. doi:10.1111/rssb.12027.
[20] J.R. Zubizarreta, Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data, Journal of the American Statistical Association. 110 (2015) 910–922. doi:10.1080/01621459.2015.1023805.
[21] K.C.G. Chan, S.C.P. Yam, Z. Zhang, Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting, J. R. Stat. Soc. B. 78 (2016) 673–700. doi:10.1111/rssb.12129.
[22] Q. Zhao, Covariate Balancing Propensity Score by Tailored Loss Functions, ArXiv:1601.05890 [Stat]. (2016). http://arxiv.org/abs/1601.05890 (accessed November 7, 2017).
[23] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological). 58 (1996) 267–288.
[24] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction., Springer, New York, NY, 2009. doi:10.1007/978-0-387-84858-7_2.
[25] A. Agresti, Foundations of linear and generalized linear models, John Wiley & Sons, 2015.
[26] T. Minka, A comparison of numerical optimizers for logistic regression, 2003.
[27] J. Pearl, Causal inference in statistics: An overview, Statist. Surv. 3 (2009) 96–146. doi:10.1214/09-SS057.
[28] L. Breiman, Random Forests, Machine Learning. 45 (2001) 5–32. doi:10.1023/A:1010933404324.
[29] V. Dorie, J. Hill, U. Shalit, M. Scott, D. Cervone, Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, ArXiv:1707.02641 [Stat]. (2017). http://arxiv.org/abs/1707.02641 (accessed November 19, 2017).

APPENDIX A: DATA PREPARATION

For this study, we used the IMS Health Surveillance Data Incorporated (SDI) medical claims databases containing anonymous, aggregated claims data of ~21 million patients from all major regions of the United States. Data consist of diagnostic records (Dx), prescription records (Rx) and hospitalization records (Hosp). SDI is constructed as a provider-centric open database, in that it collects all data from providers (e.g., pharmacy) and identifies and links patients within that data. As such, SDI may contain unknown data gaps for individual patients if they visit providers that are not covered. To increase data reliability, we only considered data stretches in which there was 80% continuous monthly eligibility (in 1-year windows) in any of the SDI pharmacy, physician, or hospital databases, and quarterly pharmacy eligibility.

This study was designed as a standard retrospective observational cohort study. The SDI data span 7 years, from January 1st, 2006, until September 31st, 2012. We defined the index date as a point in time where exactly one AED was started or added to the patient's baseline treatment regimen, i.e., it was not part of the treatment just before the index date. This added drug is considered as the exposure determining the outcome and it is classified as Older (in which case we denote E=0) or Newer (E=1). For this study, Newer AEDs were defined as Lacosamide, Levetiracetam, Oxcarbazepine, Pregabalin, Retigabine, Tiagabine, Topiramate, Felbamate, Lamotrigine, and Gabapentin, while Older AEDs were defined as Carbamazepine, Phenobarbital, Phenytoin, Primidone, and Sodium Valproate. The drug that is added is termed the index drug. More precisely: The index date is defined as the first valid treatment change event in which only one drug (from the Older and Newer AED lists above) is added or started. The conditions for an event to be a valid index date are defined as:

- The patient has at least 1 year of data before and after the index date.
- The patient has at least 3 months of Rx eligibility before index date.
- During the 1-year period post index date, the patient is on some AED for at least 50% of the days. This is designed to exclude patients who are not actively consuming AEDs.
- The treatment was unchanged during the 30-day period after the index date (to eliminate rescue medication in favor of chronic treatment).
- There cannot be a prescription for the index drug in the year pre-index date.

To capture data from patients with epilepsy rather than from patients receiving AEDs for other indications, a patient had to fulfill the below criteria to be included:

- Diagnoses criterion: At least one International Classification of Diseases, Ninth Revision (ICD-9) epilepsy diagnosis code 345.* or two seizure diagnosis codes (780.39) at any time in the data.
- Prescription criterion: At least one claim for an AED at any time in the data. This claim must be from a pharmacy with 80% stability (existence of monthly pharmacy claims data) over the entire data period.
- Overall AED criterion: throughout the patient record, the patient had to have at least one AED claim which is not gabapentin/pregabalin or there had to be at least one gabapentin/pregabalin prescription from a physician whose specialty is one of: Neurology, Clinical Neurophysiology, Child Neurology, Neurological Surgery.
- In addition, we focused on adults, requiring more than 16 years of age at beginning of data.
- To avoid patients in whom AEDs were prescribed for indications other than epilepsy, the specialty of the physician prescribing the index drug was not allowed to be related to pain management or surgery.
- Patients who met all of the above but for which no valid index date could be found, were excluded.