Detection of Anomalous Patterns Associated with the Impact of Medications on 30-Day Hospital Readmission Rates in Diabetes Care

William Ogallo, RPh, PhD, Girmaw Abebe Tadesse, PhD, Skyler Speakman, PhD, Aisha Walcott-Bryant, PhD
IBM Research – Africa, Nairobi, Kenya

Abstract

Improving quality of care in diabetes requires a good understanding of variations in diabetes outcomes and related interventions. However, little is known about the impact of diabetes interventions on outcome measures at the subpopulation-level. In this study, we developed methods that combine causal inference techniques with subset scanning techniques to study the heterogeneous effects of treatments on binary health outcomes. We analyzed a diabetes dataset consisting of 70,000 initial inpatient encounters to investigate the anomalous patterns associated with the impact of 4 anti-diabetic medication classes on 30-day readmission in diabetes. We discovered anomalous subpopulations where the likelihood of readmission was up to 1.8 times higher than that of the overall population suggesting subpopulation-level heterogeneity. Identifying such subpopulations may lead to a better understanding of the heterogeneous effects of treatments and improve targeted intervention planning.

Introduction

Thirty-day hospital readmission is an important outcome measure for assessing the quality of care given to diabetes patients. Reducing readmission rates among diabetes patients could improve care and reduce care-related costs[1]. However, although diabetes patients have an increased risk of readmission, little research has been done on this subject[1]. Some of the key barriers to understanding the risk factors of readmissions in diabetes are complicated by the natural variations in diabetes outcomes and related interventions. For example, care providers may use different treatments for their patients, and patients may respond differently to the same treatments. To improve the quality of care in diabetes care, there is a need for robust approaches for investigating variations at the subpopulation level. This is particularly useful since methods that analyze individual patients may fail to identify subtle patterns that are discernable when groups of patients are considered collectively, while methods that generate aggregate statistics for entire populations may fail to detect small-scale patterns[2].

Traditional approaches to investigating subpopulation-level heterogeneity have relied on manual stratification of covariate profiles. For example, an outcome such as mortality can be stratified by age, gender, and ethnicity to identify high-risk subpopulations. However, this is limited to analyzing only a few features beyond which it becomes computationally infeasible. Furthermore, these approaches lack a data-driven knowledge discovery aspect as investigators must suggest a priori which features they would like to stratify across, and may also inadvertently lead to data manipulation as investigators attempt to produce desired p-values (‘p-hacking’). Machine learning approaches have also been used to investigate heterogeneity. Techniques such as LASSO regression can be used to select important covariates, while decision tree regression can be used to recursively partition data. However, these approaches are either subject to several modeling assumptions and limitations or lack adequate interpretability[3]. Fortunately, recent advancements in the anomalous pattern detection literature enable the scalable and unsupervised discovery of specific subpopulations (subsets) that are anomalous. These subset scanning methods focus on identifying anomalous subsets of records in a multidimensional array that differ from expected behavior. Herein, anomalousness is quantified using a scoring function that is typically a log-likelihood ratio statistic[2]. The scoring function is maximized over the exponentially-many combinations of feature values to identify the subset with the highest score. This function must satisfy the linear time subset scanning (LTSS) property so that the search can be done in linear rather than exponential time[2].

Some of the key subset scanning techniques include Bias-Scan[4], treatment effect subset scanning (TESS)[3], and anomalous patterns of care (APC) Scan[5]. Bias-Scan focuses on the discovery of the subpopulation with the most divergence between the true outcomes and the predicted probabilities of a binary classifier. TESS discovers heterogeneous treatment effects by identifying the subpopulation in a randomized controlled trial that is most significantly impacted by the studied treatment. APC Scan extends TESS to enable anomalous pattern detection in observation data by incorporating multiple treatments and propensity score weighting to account for observable differences between treated and untreated patients. While these techniques can be applied in health and research informatics domains, certain limitations have to be addressed to improve their utility. For example, Bias-Scan is primarily used for the assessment of bias in predictive binary classifiers and although it analyzes binary outcomes, it has not been adopted for use in the assessment of heterogeneous effects of treatments. On the other hand, TESS and APC Scan are primarily designed for scalar outcomes and are currently limited to the discovery of heterogeneous effects of single interventions in randomized controlled trials (i.e., TESS) or multiple interventions with the assumption of temporal independence between the interventions (i.e., APC).

The overarching goal of our research is to extend and generalize the application of anomalous pattern detection techniques from subset scanning literature to enable the efficient discovery of anomalous patterns in large-scale observational health data such as electronic health records. The objectives of this study were threefold. First, we proposed an approach for selecting the least biased propensity score (PS) model among multiple PS models to overcome treatment selection bias in observational studies. Second, we developed algorithms combining causal inference and anomalous pattern detection techniques to discover the heterogeneous effects of interventions on binary outcomes. Third, we demonstrated the application of the developed techniques to discover anomalous patterns associated with the impact of diabetes medications on 30-day hospital readmission in diabetes care.

Methods

Propensity score based bias smoothing

Causal inference and anomalous pattern detection on observational studies require smoothing of the treatment assignment bias that accounts for observable differences (bias) between treated and untreated groups. Propensity score models, which are often used for such tasks, model the probability of receiving a treatment (p_s) conditioned on observed baseline covariates, i.e., $p_s(\mathbf{X}) = p(Z = 1|\mathbf{X})$, where X is a set of covariates for a given sample, and Z is a particular treatment assigned.

Once the propensity score for each sample in a study is computed, several propensity-based smoothing techniques could be applied[6]. These include the following: (a) *Inverse Propensity of Treatment Weighting (IPTW)* that uses weights based on propensity scores to generate synthetic samples such that the distribution of covariates is independent of the treatment; (b) *Propensity Score Matching* that matched sets of treated and untreated subjects who share a similar value of the propensity score; (c) *Stratification on the Propensity Score* that stratifies subjects into mutually exclusive subsets (e.g. quintiles) based on their propensity scores.; and (d) *Covariate Adjustment Using the Propensity Score* in which the outcome variable is regressed on an indicator variable denoting treatment status and the propensity score.

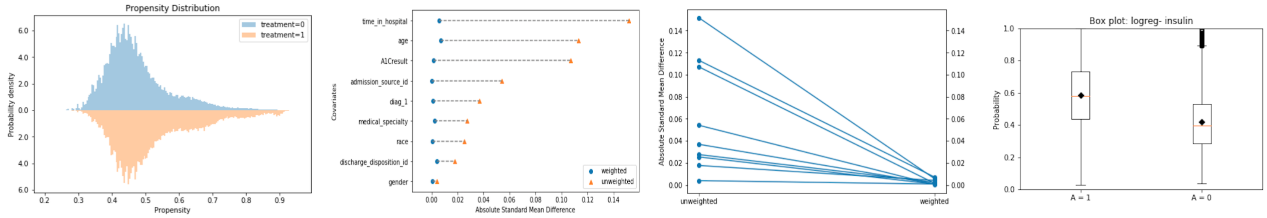
Among these approaches, we selected IPTW due to its effectiveness compared to the others as it does not require matching or stratification that might result in discarding unmatched samples or suboptimal stratification. IPTW uses weights based on propensity scores to generate synthetic samples such that the distribution of covariates is independent of the treatment. These weights include: Average Treatment Effects (w_{ATE}); Stabilized Average Treatment Effects (w_{ATE_stab}) and Average Treatment Effect on the Treated (w_{ATT}), which could be computed as follows:

$$w_{ATE} = \frac{Z}{p_s} + \frac{1-Z}{1-p_s}, \quad w_{ATE\_stab} = \frac{Zp(Z=1)}{p_s} + \frac{1-Z(p(Z=0))}{1-p_s}, \quad w_{ATT} = Z + p_s\frac{1-Z}{1-p_s}.$$

Balance diagnosis and evaluation of positivity violation

Propensity score-based bias smoothing could be achieved using different binary classification algorithms, such as logistic regression and random forest. Thus, it is important to evaluate the balance of the treatment bias achieved by the propensity model. To this end, existing balance diagnosis methods could be grouped into two: *graphical* and *quantitative* methods.

Graphical methods provide visualizations to qualitatively evaluate the balance between treated and untreated subsets. Examples of graphical balance diagnosis methods include the Propensity Score Distribution, Covariate Balance Love Plot, Covariate Balance Slope Plot, and Covariate Balance Box Plot as shown in Fig. 1. Propensity score distribution presents the probability density of treated and untreated groups across propensity score. Overlapping between the two density functions suggests balancing where a horizontal deviation of these distributions signals the lack of balancing and hence positivity assumption violations. Covariate Balance Love Plot provides the absolute standard mean difference for each covariate between the treated and untreated group, before and after weighting is applied. A smaller deviation (< 0.1) signals good balancing. The Covariate Balance Slope Plot is another visualization of the absolute mean differences and an alternative to the Covariate Balance Love Plot. Finally, the Covariate Balance Box Plot shows the mean propensity score for the treated and untreated groups, where good balance is depicted from similar mean propensity scores of the two groups.



(a) propensity score distribution **(b)** covariate balance love plot **(c)** covariate balance slope plot **(d)** covariate balance box plot

Figure 1: Examples of graphical methods to assess the balance of observed covariates in treated and comparison groups using propensity scores and inverse probability of treatment weighting.

Quantitative methods provide quantifiable values regarding the balancing by the propensity score-based weights, which could then also help to evaluate positivity assumption violations[7]. Examples of quantitative methods include the *standardised difference* (s_d)[7], the *Kolmogorov-Smirnov test statistic* (k_s)[8], and *overlapping index* (o_i)[9]. Given a covariate profile, x, which becomes x_t for treated group and x_c for the comparison group, these quantitative values could be obtained as follows:

$$s_d(x) = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{s_{xt}^2 + s_{xc}^2}}, \qquad k_s = Max|p_{st}(x) - p_{sc}(x)|, \qquad o_i = \int (min[p_{st}(x), p_{sc}(x)]dx)$$

where $\bar{x}_t$ and $\bar{x}_c$ represent mean of treated (x_t) and comparison (x_c) groups, respectively; s_{xt} and s_{xc} represent the standard deviation of x_t and x_c, respectively. Similarly, $p_{st}x$ and $p_{sc}x$ represent the propensity score for x_t and x_c, respectively.

Unified Performance Index

The quantitative approaches used to diagnose both the imbalance of the treated and comparison groups as well as positivity assumption violations are often used separately, yet it is important to have a unified evaluation framework that takes into consideration the different quantitative evaluation techniques. To this end, we introduced the Unified Performance Index (UPI) that takes into account the overlapping index, mean stabilized ATE weights, KS statistic, and mean of the weighted standardized mean difference. Better balancing and lesser positivity assumption violation diagnosis correspond to a higher UPI score. The UPI maximizes the overlapping index o_i, minimizes the absolute mean standardized differences for covariates s_d, minimizes the KS test statistic k_s and minimizes the deviation of mean stabilized ATE weights w_{ATE_stab} from unit value. Thus, UPI is formulated to reflect these proportionality characteristics as follows:

$$UPI = \frac{o_i}{s_d + k_s + abs(1 - w_{ATE\_stab})}$$

Here, o_i has a domain of $[0, 1]$ such that $o_i = 0$ indicates that the propensity score distributions among the treated and untreated groups are completely separated. Conversely $o_i = 1$ indicates that the two distributions the same. The s_d has a domain of $[0, 1]$ such that $s_d = 0$ implies that there is no difference in the mean or prevalence of variables between the treated and untreated groups. k_s also has a domain of $[0, 1]$ such that $k_s = 0$ if the cumulative distributions of propensity scores among the treated and untreated groups are identical, and $k_s = 1$ if the distributions are completely distinct. Lastly, w_{ATE_stab} has a domain of $[-\infty, +\infty]$ such that mean stabilized weights that are further from one are indicative of higher degrees of the violation of the positivity assumption.

Subset scanning for anomalous pattern detection

In subset scanning literature, the pattern detection problem can be framed as a search over all subsets in a multidimensional array that spans any combination of feature values to identify the most anomalous subset, i.e. the subset with the most evidence of divergence from expected behavior. Scanning is achieved by maximizing a scoring function, $F(S)$, over all subsets to identify the highest-scoring subset, $S^* = \arg\max_S F(S)$. This approach can, therefore, be used to reveal hidden anomalous subsets that may not be obvious when inspecting individual features manually. The scoring functions exploit a mathematical property, the Linear Time Subset Scanning (LTSS) property[2], which proves that the values of a given discrete/discretized feature can be ordered optimally using a priority function such that scanning is done without requiring an exhaustive search and is guaranteed to be completed in linear time ($O(n)$) rather than in exponential time ($O(2^n)$).

As previously highlighted, several subset scanning techniques have been developed. In this study, we specifically extend the Bias-Scan methodology. The goal of Bias-Scan is to discover the subpopulation with the most divergence between the true outcomes and the predicted probabilities of a binary classifier[4]. Given tabular data with discrete/discretized covariates, a binary outcome, y_i, and predictions generated by a binary classifier, $\hat{p_i}$, Bias-Scan maximizes a Bernoulli likelihood ratio scoring statistic, $score_{bias}(S)$, that quantifies bias in a given subgroup. The algorithm identifies the subgroup that has the most evidence of having the expected odds differing from the predicted odds. Here, the null hypothesis is that prediction odds are correct across all subgroups, $H_0 : odds(y_i) = \frac{\hat{p_i}}{1-\hat{p_i}}$; while the alternative hypothesis assumes a constant multiplicative increase in the prediction odds for some given subgroup, $H_1 : odds(y_i) = q\frac{\hat{p_i}}{1-\hat{p_i}}$ where $q > 1$. The scoring function in Bias-Scan is:

$$score_{bias}(S) = \max_q log(q) \sum_{i \in S} y_i - \sum_{i \in S} log(1 - \hat{p_i} + q\hat{p_i})$$

Consequently, subsets in which records have larger numbers of $y_i = 1$ but smaller corresponding p_i will have higher scores. To detect the anomalous subgroups, Bias-Scan uses the Multi-Dimensional Subset Scanning (MDSS) algorithm[10]. Given a multi-dimensional array of discrete/discretized features, MDSS optimizes Bias-Scan's likelihood ratio statistic over all subsets of values of each feature conditioned on the current subset of all other features in the multi-dimensional array. To do so efficiently and exactly, MDSS satisfies the LTSS property[2] with a priority function computed as the ratio of the observed odds and the expected odds. This priority function[3] ranks the values of a given feature and then select the highest-scoring subset as the subset consisting of the "top-k" priority values for some $k \in [1, \ldots, J]$. MDSS iterates over all features in the multidimensional array until convergence to a local maximum is found. The global maximum is subsequently optimized using multiple random restarts.

Anomalous subgroup detection for binary outcomes

Our study combines propensity score techniques with the Bias-Scan to discover anomalous patterns associated with medication classes used in diabetes care. Here, we specifically proposed three algorithms: Conditional Automated Stratification Scan (CASS), Matched Conditional Automated Stratification Scan (mCASS), and Weighted Conditional Automated Stratification Scan (wCASS). These algorithms analyze tabular data with discrete/discretized covariate profiles X, a single binary treatment $Z \in \{0, 1\}$, and a single binary outcome $Y \in \{0, 1\}$. The key steps in the algorithms are described in Table 1 and discussed in detail below.

1. Conditional Automated Stratification Scan (CASS)

The CASS algorithm represents our simplest extension of the Bias-Scan methodology to enable the estimation of heterogeneous effects of interventions on a binary outcome. In CASS, the anomalousness of any given subpopulation S of treated subjects is quantified as $E[Y_i(1) = 1|X_i \in S] < E[Y_i(0) = 1]$ for under-risked subpopulations (i.e.

Table 1: Algorithms for anomalous subgroup detection for binary outcomes

Conditional Automated Stratification Scan (CASS)	**Matched Conditional Automated Stratification Scan (mCASS)**	**Weighted Conditional Automated Stratification Scan (wCASS)**
1. Get treatment group data $Data\|_{Z=1}$ 2. Get comparison group data $Data\|_{Z=0}$ 3. For each subject i in $Data\|_{Z=1}$, estimate the counterfactual outcome as mean outcome in $Data\|_{Z=0}$, i.e. $\hat{Y}_i = E[Y(0) = 1]$ 4. Apply Bias-Scan(X, Y, $\hat{Y}$) using $Data\|_{Z=1}$ 5. Estimate statistical significance of identified subpopulation using boot-strapped randomization testing	1. Get the treatment's propensity scores from the best propensity score model 2. Get the treatment's logit of the propensity score 3. Get treatment group data $Data\|_{Z=1}$ 4. Get comparison group data $Data\|_{Z=0}$ 5. For each subject i in the $Data\|_{Z=1}$ (a) Identify nearest neighbors in $Data\|_{Z=0}$ as those within 0.2SD of the logit of the propensity score (b) Estimate the counterfactual outcome $\hat{Y}_i$, as the average outcome among the identified nearest neighbors 6. Apply Bias-Scan(X, Y, $\hat{Y}$) using $Data\|_{Z=1}$ 7. Estimate statistical significance of identified subpopulation using boot-strapped randomization testing	1. Get the treatment's propensity scores from the best propensity score model 2. Compute the average treatment effect on the treated (ATT) weights (w_{ATT}) 3. Get treatment group data $Data\|_{Z=1}$ 4. Get comparison group data $Data\|_{Z=0}$ 5. For each subject i in $Data\|_{Z=1}$, estimate the counterfactual outcome $\hat{Y}_i$ as the ATT-weighted mean expected outcome in $Data\|_{Z=0}$ 6. Apply Bias-Scan(X, Y, $\hat{Y}$) using $Data\|_{Z=1}$ 7. Estimate statistical significance of identified subpopulation using boot-strapped randomization testing

lower than expected outcomes), and $E[Y_i(1) = 1|X_i \in S] > E[Y_i(0) = 1]$ for over-risked subpopulations (i.e. higher than expected outcomes). Here, $Y_i(1)$ denotes the occurrence the an outcome for a treated subject, $E[Y_i(1) = 1|X_i \in S] = \frac{1}{N_1}\sum_{i=1 i\in S}^{N_1} Y_i$ is the probability of the outcome in the subpopulation S, and $E[Y_i(0) = 1] = \frac{1}{N_0}\sum_{i=1}^{N_0} Y_i$ is the marginal probability of occurrence of the outcome among the comparison group subjects. Therefore, CASS assumes that the counterfactual outcome for each treated subject is $E[Y(0) = 1]$ and that the average unit-level causal effect of the treatment for each treated subject is $Y(1)_i - E[Y(0) = 1]$. CASS then searches for a specific most anomalous subpopulation S^* as the subpopulation in which the probability of the outcome conditioned on belonging to this subpopulation has the most evidence of being divergent from the marginal probability of the outcome among all the comparison group subjects. Procedurally, the key steps in CASS are described in Table 1.

2. Matched Conditional Automated Stratification Scan (mCASS)

The mCASS algorithm combines propensity score matching with Bias-Scan to discover heterogeneous treatment effects across subpopulations. In mCASS, the counterfactual outcome for each treated subject is determined by propensity score matching between pairs of treated and comparison group subjects who have similar propensity scores. Consequently, in mCASS, the average treatment effect is estimated as the average of the differences between the actual versus counterfactual outcome within each pair. While several propensity score matching techniques can be used in mCASS, we specifically use the nearest neighbor caliper matching[6]. In this approach, matching is done on the logit of the propensity scores using calipers of width 0.2 standard deviation of the logit of the propensity score as a threshold for matching[6]. The key steps in mCASS are described in Table 1.

3. Weighted Conditional Automated Stratification Scan (wCASS)

The wCASS algorithm uses the IPTW derived from propensity scores. Specifically we use the previously described average treatment effect on the treated (ATT) weights, w_{ATT}. When using wCASS, the anomalousness of a subpopulation S is quantified as $E[w_iY_i(1)|X_i \in S] < E[w_iY_i(0)]$ for under-risked subpopulations and as $E[Y_i(1)|X_i \in S] > E[Y_i(0)]$ for over-risked subpopulations. Here, $w_iY_i(1)$ denotes the weighted outcome for a treated subject, $E[w_iY_i(1)|X_i \in S] = \frac{1}{N_1}\sum_{i=1 i\in S}^{N_1} w_iY_i$ is the weighted probability of the outcome in the subpopulation S, and

$E[w_iY_i(0)] = \frac{1}{N_0}\sum_{i=1}^{N_0} w_iY_i$ is the weighted marginal probability of occurrence of the outcome among the comparison group subjects. Procedurally, wCASS works as described in Table 1.

Randomization Testing

All the 3 algorithms, CASS, mCASS, and wCASS use randomization testing to estimate the statistical significance of the detected anomalous subgroups. To do so, we draw 1000 bootstrapped random subpopulations and compute the subset score for each subpopulation drawn, and then compare each score to the score for the detected anomalous subpopulation. We compute the empirical p-value as $\frac{(r+1)}{(n+1)}$ where r is the number of scores greater than or equal to that actual score and n is the total randomization scores for bootstrapped samples.

Experiment and Results

To validate the algorithms we developed, we used a case study whose goal was to detect the anomalous patterns associated with the impact of medications on 30-day hospital readmission in diabetes. Figure 2 illustrates the pipeline used in the study.

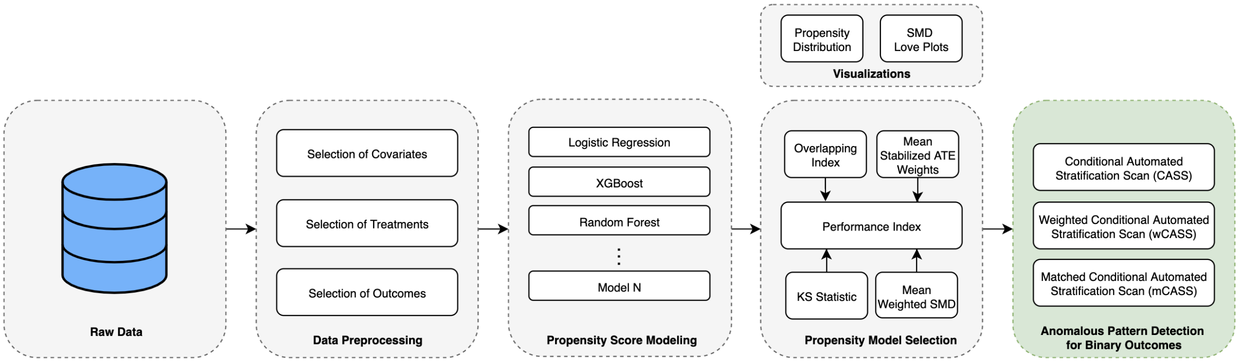

Propensity Distribution
SMD Love Plots
Visualizations
Raw Data
Selection of Covariates
Selection of Treatments
Selection of Outcomes
Data Preprocessing
Logistic Regression
XGBoost
Random Forest
⋮
Model N
Propensity Score Modeling
Overlapping Index
Mean Stabilized ATE Weights
Performance Index
KS Statistic
Mean Weighted SMD
Propensity Model Selection
Conditional Automated Stratification Scan (CASS)
Weighted Conditional Automated Stratification Scan (wCASS)
Matched Conditional Automated Stratification Scan (mCASS)
Anomalous Pattern Detection for Binary Outcomes

Figure 2: Pipeline for anomalous pattern detection associated with binary Outcomes in healthcare

The Diabetes Cohort and Dataset

We used a de-identified diabetes dataset extracted from the Health Facts database (Cerner Corporation, Kansas City, MO). The study population consisted of 69,990 patients with inclusion criteria defined as follows: (1) having an initial inpatient encounter (a hospitalization), (2) having a diagnosis of diabetes made at the initial encounter, (3) the length of stay at least 1 day and at most 14 days, (4) laboratory tests were performed during the encounter, and (5) medications were administered during the encounter[11]. The data preprocessing is described in more detail by Strack et al.[11] who also made it publicly available as supplementary material accessible at `http://dx.doi.org/10.1155/2014/781670`. We conducted additional preprocessing to generate a final dataset that consisted of a binary outcome (Readmission within 30 days), 10 discrete/discretized covariates (race, gender, age group, admission type, discharge disposition, admission source, primary diagnosis, secondary diagnosis, tertiary diagnosis, and HbA1C), and 4 intervention drug classes (Biguanides, Insulin, Sulfonylureas, and Thiazolidinediones). These preprocessing steps included dropping features with a large proportion of missing values, remapping categorical features e.g., age groups and race, mapping specific ICD9-CM codes to more general ICD9-CM codes, mapping treatment interventions to binary pharmacological classes, and mapping the readmission outcome to a binary variable. Table 2 describes the features and the distribution of the feature values in the final dataset.

Propensity Score Modeling

For each of the 4 treatment classes (biguanides, insulins, sulfonylureas, and thiazolidinediones) in the final dataset, we trained 3 propensity score models for predicting the likelihood of a patient subject receiving the treatment given

Table 2: Distribution of feature values in the final dataset

Feature	Feature Value	Overall (%)	Readmitted (%)	Not Readmitted (%)
n		69990	6285 (9.0)	63705 (91.0%)
Age	0-29	1808 (2.6)	112 (1.8)	1696 (2.7)
	30-59	21871 (31.2)	1574 (25.0)	20297 (31.9)
	60-99	46311 (66.2)	4599 (73.2)	41712 (65.5)
Gender	Female	37239 (53.2)	3365 (53.5)	33874 (53.2)
	Male	32751 (46.8)	2920 (46.5)	29831 (46.8)
Race	AfricanAmerican	12627 (18.0)	1095 (17.4)	11532 (18.1)
	Caucasian	52305 (74.7)	4807 (76.5)	47498 (74.6)
	Missing	1919 (2.7)	141 (2.2)	1778 (2.8)
	Other	3139 (4.5)	242 (3.9)	2897 (4.5)
Primary diagnosis	Circulatory	21390 (30.6)	2070 (32.9)	19320 (30.3)
	Diabetes	5748 (8.2)	524 (8.3)	5224 (8.2)
	Digestive	6488 (9.3)	520 (8.3)	5968 (9.4)
	Genitourinary	3441 (4.9)	309 (4.9)	3132 (4.9)
	Injury	4696 (6.7)	507 (8.1)	4189 (6.6)
	Musculoskeletal	4064 (5.8)	341 (5.4)	3723 (5.8)
	Neoplasm	2538 (3.6)	230 (3.7)	2308 (3.6)
	Respiratory	9491 (13.6)	693 (11.0)	8798 (13.8)
	Other	12134 (17.3)	1091 (17.4)	11043 (17.3)
Specialty of admitting physician	Cardiology	4208 (6.0)	303 (4.8)	3905 (6.1)
	Family/GeneralPractice	4978 (7.1)	485 (7.7)	4493 (7.1)
	InternalMedicine	10641 (15.2)	1039 (16.5)	9602 (15.1)
	Surgery	3751 (5.4)	297 (4.7)	3454 (5.4)
	Other	12758 (18.2)	1051 (16.7)	11707 (18.4)
	Missing/Unknown	33654 (48.1)	3110 (49.5)	30544 (47.9)
Admission source	Emergency Room	37273 (53.3)	3452 (54.9)	33821 (53.1)
	Referral	22793 (32.6)	1973 (31.4)	20820 (32.7)
	Other	9924 (14.2)	860 (13.7)	9064 (14.2)
Discharge disposition	Home	44322 (63.3)	3079 (49.0)	41243 (64.7)
	Other	25668 (36.7)	3206 (51.0)	22462 (35.3)
A1Cresult	Normal	3741 (5.3)	323 (5.1)	3418 (5.4)
	7 to 8	2866 (4.1)	247 (3.9)	2619 (4.1)
	>8	6239 (8.9)	509 (8.1)	5730 (9.0)
	No Test	57144 (81.6)	5206 (82.8)	51938 (81.5)
Time in hospital, mean (SD)		4.3 (2.9)	4.8 (3.1)	4.2 (2.9)
Time in Hospital (Categorical)	<=3 days	35146 (50.2)	2647 (42.1)	32499 (51.0)
	>3 days	34844 (49.8)	3638 (57.9)	31206 (49.0)
Biguanides	0	54628 (78.1)	5009 (79.7)	49619 (77.9)
	1	15362 (21.9)	1276 (20.3)	14086 (22.1)
Insulins	0	34268 (49.0)	2843 (45.2)	31425 (49.3)
	1	35722 (51.0)	3442 (54.8)	32280 (50.7)
Sulfonylureas	0	49228 (70.3)	4336 (69.0)	44892 (70.5)
	1	20762 (29.7)	1949 (31.0)	18813 (29.5)
Thiazolidinedione	0	60092 (85.9)	5416 (86.2)	54676 (85.8)
	1	9898 (14.1)	869 (13.8)	9029 (14.2)

his/her covariate profile. The predictor variables in each model consisted of race, gender, age, discharge disposition, admission source, time in hospital (continuous), primary diagnosis, hbA1C result, and the medical specialty of the admitting physician. Each response variable was a binary treatment class. These variables are described in Table 2. The trained propensity score models included a logistic regression model, a gradient boosting decision tree model (XGBoost), and a random forest model. Table 3 describes the modeling performance results. We note that, based on the UPI score, XGBoost consistently performed well across the treatment classes considered. Logistic regression also performs relatively well, while the random forest model performed worst across all the treatments despite having relatively better Area Under Curve results for training and test sets. The overlapping indexes for the propensity scores generated from the random forest model were lower than those of the other two models. These findings confirm a known observation that the best propensity score models are not necessarily those that are good at prediction, but those that provide better overlapping between the propensity scores of the treated versus untreated subjects.

Characteristics of the most anomalous subpopulations discovered

Table 4 shows the anomalous pattern detection results for the different treatments analyzed and algorithms used. Each algorithm was able to identify heterogeneous treatment effects by discovering the most anomalous subgroup

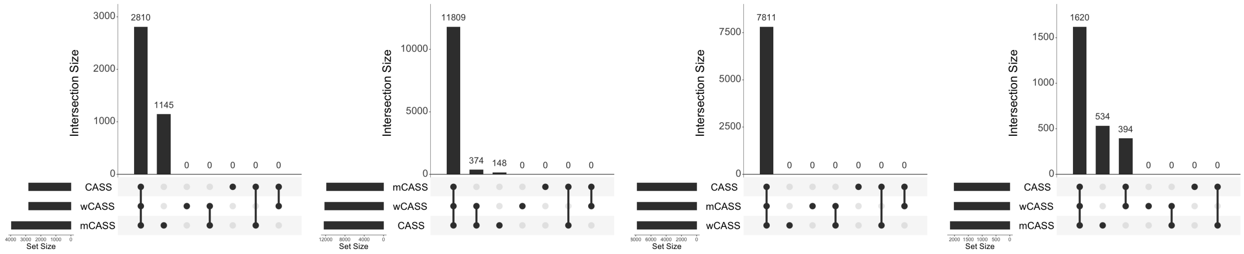
Table 3: Propensity Score Modeling Performance Results

Treatment	Model	Prob of Treatment	Train AUC	Test AUC	Mean Std Diff Unweighted	Mean Std Diff Weighted	KS Test Statistic	KS Test p-value	Overlapping Index	Mean ATE Stab Weight	UPI Score
Biguanides	XGBoost	0.219	0.626	0.616	0.218	0.015	0.175	<0.001	0.707	0.998	3.692
	Logistic Regression	0.219	0.606	0.607	0.218	0.014	0.152	<0.001	0.743	1.319	1.529
	Random Forest	0.219	0.691	0.615	0.218	0.039	0.24	<0.001	0.616	1.271	1.12
Insulins	Logistic Regression	0.51	0.624	0.62	2.533	0.014	0.172	<0.001	0.708	1.001	3.787
	XGBoost	0.51	0.663	0.642	2.533	0.015	0.221	<0.001	0.64	0.995	2.65
	Random Forest	0.51	0.679	0.639	2.533	0.05	0.241	<0.001	0.613	0.974	1.938
Sulfonylureas	XGBoost	0.297	0.613	0.607	1.099	0.013	0.158	<0.001	0.73	0.998	4.218
	Logistic Regression	0.297	0.605	0.606	1.099	0.008	0.15	<0.001	0.741	1.166	2.284
	Random Forest	0.297	0.67	0.601	1.099	0.035	0.223	<0.001	0.637	1.137	1.615
Thiazolidinediones	XGBoost	0.141	0.598	0.57	0.362	0.027	0.128	<0.001	0.778	0.998	4.962
	Logistic Regression	0.141	0.578	0.571	0.362	0.01	0.105	<0.001	0.815	1.514	1.295
	Random Forest	0.141	0.689	0.569	0.362	0.035	0.233	<0.001	0.625	1.454	0.866

conditioned on patients receiving a given treatment. This heterogeneity is described in terms of risk difference, relative risk, and odds ratio in the overall population of treated subjects versus the most anomalous subpopulation identified. We observe that for each treatment, the discovered anomalous subpopulations are relatively similar in sizes across the CASS, mCASS, and wCASS algorithms. We also observe that for each treatment, the algorithms resulted in similar measures of effect across the identified anomalous subpopulations. This observation could be explained by the high degrees of overlaps in the subpopulations identified by the different algorithms as illustrated in Figure 1.

Table 4: Anomalous Pattern Detection Results

		Anomalous Subpopulation			Risk Difference		Relative Risk		Odds Ratio	
Treatment	**Algorithm**	**Size**	**Score**	**P-Value**	**Population**	**Subpopulation**	**Population**	**Subpopulation**	**Population**	**Subpopulation**
Biguanides	CASS	2810 (18.3%)	39.1	0.004	0	0.1	0.9	1.6	0.9	1.7
	mCASS	3955 (25.7%)	41.1	0.003	0	0	0.9	1.5	0.9	1.6
	wCASS	2810 (18.3%)	44	0.003	0	0.1	0.9	1.6	0.9	1.7
Insulins	CASS	12331 (34.5%)	180.5	0.002	0	0.1	1.2	1.6	1.2	1.7
	mCASS	11809 (33.1%)	169.5	0.002	0	0.1	1.1	1.6	1.2	1.7
	wCASS	12183 (34.1%)	158.8	0.002	0	0	1.1	1.6	1.1	1.7
Sulfonylureas	CASS	7811 (37.6%)	79.3	0.002	0	0	1.1	1.5	1.1	1.6
	mCASS	7811 (37.6%)	61.7	0.002	0	0	1	1.4	1	1.5
	wCASS	7811 (37.6%)	65.7	0.002	0	0	1	1.4	1	1.5
Thiazolidinediones	CASS	2014 (20.3%)	37.9	0.003	0	0.1	1	1.7	1	1.8
	mCASS	2154 (21.8%)	34.5	0.002	0	0.1	1	1.6	1	1.7
	wCASS	2014 (20.3%)	39.1	0.003	0	0.1	1	1.7	1	1.8



(a) Biguanides **(b)** Insulins **(c)** Sulfonylureas **(d)** Thiazolidinediones

Figure 3: Overlaps between anomalous subpopulations identified by CASS, mCASS, and wCASS

Interestingly, the subgroups with the largest measures of effect pertained to the use of thiazolidinediones, suggesting a stronger heterogeneous treatment effect of this class of drugs on 30-day hospital readmission. By way of example, the subsets discovered by CASS and wCASS for thiazolidinediones were identical and suggests that patients who use thiazolidinediones and are aged 60 years or older; and are Caucasian or African American or have their race information missing; and have a primary diagnosis that is not musculoskeletal; and were admitted by specialists who are not cardiologists; and had HbA1C >8 or had no HbA1C test conducted at the time of admission; and were discharged to destinations other than their homes, were 1.8 times more likely to be readmitted within 30 days than the average non-treated population. This subpopulation had a 30-day readmission rate that differed the most from the expected 30-day readmission rate determined as the global mean among untreated subjects.

Discussion

This study aimed at developing and demonstrating the application of techniques for assessing the causal heterogeneous effects of binary interventions on binary outcomes in observational health data such as electronic health records. To this end, the study proposed a unified performance index (UPI) for choosing the best propensity score model among multiple propensity score models and describes how algorithms from the causal inference and anomalous pattern detection literature could be leveraged to discover anomalous patterns of care. Furthermore, the demonstrates how the developed algorithms can be used to detect the heterogeneous treatment effects captured in electronic health records.

We discovered highly similar subpopulations among which the use of anti-diabetic medication classes (biguanides, insulins, sulfonylureas, and thiazolidinediones) was associated with an increased likelihood of being readmitted within 30 days after the index inpatient admission. Interestingly, thiazolidinedione therapy has previously been associated with a higher risk of hospital readmissions on average[12–14]. The findings from our study suggest that certain subpopulations may be differentially affected by this class of drugs as well as other commonly used classes of anti-diabetic medication. However, we take cognizance of the fact that our approach should be viewed as a method for generating hypotheses about the specific subpopulations that are most likely to be impacted by the interventions. How such heterogeneity occurs or is realized is beyond the scope of the current approach and further investigations are warranted to confirm the generated hypotheses.

Current literature has primarily focused on studying the average treatment effect of interventions on binary outcomes[6], or on studying heterogeneous treatment effects of single interventions in clinical trials[3]. These approaches are, however, done separately. To the best of our knowledge, our study is the first to leverage and extend both causal inference and subset scanning techniques to study the effect of interventions on binary health outcomes. The techniques we have developed can provide researchers, care providers, and other stakeholders with the ability to identify positive or adverse clinical practices and subsequently institute better care delivery plans based on the identified insights. They can also be used as a basis for risk analysis and recommendation of follow-up interventions to improve care experiences and outcomes for individual patients, especially in differential service delivery and targeted intervention planning settings. Furthermore, they can enable payers to identify drivers of poor outcomes and unnecessary costs across patient subpopulations.

Whereas we took the necessary steps to ensure robustness in our study, several limitations can be observed. First, as would be expected, feature selection and engineering can affect propensity scoring modeling and the anomalous pattern detection, subsequently bias findings. We minimized this by testing our approach on a diabetes dataset validated by Strack et al.[11] while maintaining the features and feature values used in the aforementioned study. Second, the UPI score is currently unbounded and ranges from zero to infinity with higher scores implying better performance. However, a monotonic function that maps the UPI score to [0,1] can be applied without loss of generality. As part of our future work, we intend to refine, test, and compare different formulations of the UPI across different propensity score models trained on several publicly available datasets. Third, the subset scanning approach applied in this study can be misconstrued as conducting multiple hypotheses testings. However, we consider this and use parametric bootstrapped randomization tests to determine the statistical validity of anomalous subpopulations discovered by our algorithms. Fourth, the results of the subset scanning process can be complex and difficult to interpret even for persons with domain expertise. As part of our future work, we intend to incorporate into our pipeline a penalized version of Bias-Scan that uses a penalty function to minimizes the complexity and maximize the size of the identified subpopulations[15]. Lastly, our current approach can only analyze binary interventions and binary outcomes. We, however, acknowledge that there are different healthcare outcomes that can be binary (e.g. mortality), discrete (e.g. number of days spent in hospitals), continuous (e.g. blood glucose levels). At the same time, the intervention space in healthcare is often complex and can range from single interventions given only once (e.g. single dose medications or vaccinations), to multiple interventions used simultaneously (e.g. drug combinations in regimens), to sequential interventions (e.g. temporally dependent drug regimens, clinical pathways). As part of our future work, we plan to develop techniques for discovering anomalous patterns associated with these different types of interventions and outcomes in healthcare data. Additionally, we intend to compare our approach to state-of-the-art subgroup analysis techniques and to generate meaningful clinical insights, perspectives, and interpretations of the discovered anomalous subpopulation through counterfactual analyses of modifiable risk factors that could serve as a baseline for targeted intervention planning.

Conclusion

We studied 30-day readmission in diabetes using methods that combine techniques from causal inference and anomalous pattern detection literature to study the heterogeneous effects of treatments. This study shows that a unified performance index can be used to select the best propensity score models among multiple models. It also shows that techniques such as propensity score matching and inverse probability of treatment weighting could be leveraged to study the impact of binary treatment on binary outcomes at the subpopulation-level and in a disciplined statistical approach. Furthermore, the study shows that for a given treatment, the studied algorithms result in the identification of subpopulations that are highly similar in terms of common covariate characteristics. Lastly, the study demonstrates that for certain subpopulations, the likelihood of 30-day hospital readmission among index diabetes encounter may be differentially impacted by the use of some anti-diabetic medication classes and that these differential effects may not be discernible at the overall populations. Future work includes delineating the merits and demerits of our algorithms, penalizing complexity while maximizing subset sizes for easier interpretability, and generalizing the approaches for application across disparate intervention and outcome data types.

References

[1] Daniel J Rubin. Correction to: hospital readmission of patients with diabetes. *Current diabetes reports*, 18(4):21, 2018.

[2] Daniel B Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

[3] Edward McFowland III, Sriram Somanchi, and Daniel B Neill. Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *arXiv preprint arXiv:1803.09159*, 2018.

[4] Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.

[5] Edward Somanchi, Sriram McFowland III and Daniel B Neill. Detecting anomalous patterns of care using health insurance claims. *Presented at Conference on Information Systems and Technology*, 2017.

[6] Peter C Austin and Elizabeth A Stuart. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, 26(6):2505–2525, 2017.

[7] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

[8] Jasjeet S Sekhon. Alternative balance metrics for bias reduction in matching methods for causal inference. *Survey Research Center, University of California, Berkeley*, 2007.

[9] Massimiliano Pastore and Antonio Calcagnì. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in psychology*, 10:1089, 2019.

[10] Daniel B Neill, Edward McFowland III, and Huanian Zheng. Fast subset scan for multivariate event detection. *Statistics in medicine*, 32(13):2185–2208, 2013.

[11] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[12] Frederick A Masoudi, Silvio E Inzucchi, Yongfei Wang, Edward P Havranek, JoAnne M Foody, and Harlan M Krumholz. Thiazolidinediones, metformin, and outcomes in older patients with diabetes and heart failure: an observational study. *Circulation*, 111(5):583–590, 2005.

[13] Fei-Yuan Hsiao, Yi-Wen Tsai, Yu-Wen Wen, Pei-Fen Chen, Hao-Yu Chou, Chen-Huan Chen, Ken N Kuo, and Weng-Foung Huang. Relationship between cumulative dose of thiazolidinediones and clinical outcomes in type 2 diabetic patients with history of heart failure: a population-based cohort study in taiwan. *Pharmacoepidemiology and drug safety*, 19(8):786–791, 2010.

[14] Silvio E Inzucchi, Frederick A Masoudi, Yongfei Wang, Mikhail Kosiborod, Joanne M Foody, John F Setaro, Edward P Havranek, and Harlan M Krumholz. Insulin-sensitizing antihyperglycemic drugs and mortality after acute myocardial infarction: insights from the national heart care project. *Diabetes Care*, 28(7):1680–1689, 2005.

[15] Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404, 2016.