Veridical Causal Inference for Comparative Effectiveness Research Using Medical Claims

Authors: Ryan D. Ross, MS[1]
Xu Shi, PhD[1]
Megan E. V. Caram, MD, MS[2,3,4]
Pheobe A. Tsao, MD[2]
Paul Lin, MS[4]
Amy Bohnert, PhD[3,4,5]
Min Zhang, PhD[1]
Bhramar Mukherjee, PhD[1]

Author Affiliations:
[1]Department of Biostatistics, School of Public Health, University of Michigan
[2]Department of Internal Medicine, Division of Hematology/Oncology, University of Michigan Medical School
[3]VA Health Services Research & Development, Center for Clinical Management and Research, VA Ann Arbor Healthcare System, Ann Arbor, Michigan
[4]Institute for Health Policy and Innovation, University of Michigan Medical School
[5]Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI;

Corresponding Author Contact Information:
Ryan D. Ross
University of Michigan: School of Public Health
1415 Washington Heights
Ann Arbor, MI 48109
Phone: (775) 688-9091
Email: rydaro@umich.edu

Acknowledgement of Research Support: Dr. Mukherjee is funded by the University of Michigan Comprehensive Cancer Center Core Grant - P30 CA 046592.
Dr. Caram is funded by a Prostate Cancer Foundation Young Investigator Award

Conflicts of Interest: The authors have no competing interests and nothing to disclose.

Key Words: insurance claims, reproducibility, propensity score, veridical data science, sensitivity analysis

Abstract

Medical insurance claims are becoming increasingly common data sources to answer a variety of questions in biomedical research. Although comprehensive in terms of longitudinal characterization of disease development and progression for a potentially large number of patients, population-based studies using these datasets require thoughtful modification to sample selection and analytic strategies, relative to other types of studies. Along with complex selection bias and missing data issues, claims-based studies are purely observational, which limits effective understanding and characterization of the treatment differences between groups being compared. All these issues contribute to a crisis in reproducibility and replication of comparative findings. This paper offers some practical guidance to the full analytical process, demonstrates methods for estimating causal treatment effects on several types of outcomes common to such studies, such as binary, count, time to event and longitudinally varying repeated measures outcomes, and aims to increase transparency and reproducibility. We provide an online version of the paper with readily implementable code for the entire analysis pipeline to serve as a guided tutorial for practitioners. The online version can be accessed at https://rydaro.github.io/. The analytic pipeline is illustrated using a sub-cohort of patients with advanced prostate cancer from the large Clinformatics TM Data Mart Database (OptumInsight, Eden Prairie, Minnesota), consisting of 73 million distinct private payer insurees from 2001-2016.

Introduction and Background

Health service billing data can be used to answer many clinical and epidemiological questions using a large number of patients and has the potential to capture patterns in health care practice that take place in the real world.[1,2,3,4,5] Such large datasets allow investigators to conduct scientific queries which may be difficult, if not practically impossible, to answer via a randomized clinical trial. For example, comparing multiple treatments that are produced by different drug companies and with varying guidelines for their use for a disease may only be feasible in a real healthcare database.[6,7] Although these large data sources offer a wealth of information, there are many challenges and drawbacks, such as confounding, selection bias, heterogeneity, missing values, duplicate records and misclassification of disease and exposures. These added complexities of these observational datasets contribute to the challenge of reproducing findings from studies using administrative health data. As regulatory agencies and pharmaceutical companies increasingly consider studying the real world evidence present in such databases, the importance of proper methodology, reporting, and reproducibility of the analysis for a broad audience of researchers is of necessity.[8,9,10,11,12,13,14] We emulate newly introduced principles from the predictability, computability, and stability (PCS) framework for veridical data science[15] to examine comparative effectiveness research questions that administrative claims data can be used to address. We provide documentation and code in R Markdown for each stage of analysis online at https://rydaro.github.io/ .

Challenges to Characterizing Treatments using Claims Data

Healthcare claims data has been extensively criticized for its use in epidemiological research.[16,17] These types of data are prone to issues such as misclassification, missing data, and bias. For example, ICD codes are entered by the care provider, and thus certain diagnoses may be missed or may not be accurate or may differ across providers.[17] Further, coding schema can change over time, such as the change from ICD-9 to ICD-10, which can further complicate analyses. Outcomes can be particularly difficult to define and identify. For example, there is no agreed upon algorithm for identifying Emergency Room visits, and thus many definitions are used.[18] While not as clean as gold standard clinical trial data, these datasets are still the best source of data for a wide variety of questions regarding drug utilization, effectiveness, and monitoring of adverse events.[19,20] Claims data have the benefit of reflecting how medications are actually being prescribed, and thus may provide a more accurate depiction of treatment benefit in practice or real-life evidence. Further, these datasets provide a wide breadth of a patient’s interaction with the healthcare system that electronic medical record (EHR) data alone,[21] going beyond just visits by adding procedures, tests, and pharmacy fills. With proper study design and methodological considerations, many of the common issues and concerns can be addressed,[8,9,10,11,12,13,14] and these large databases of longitudinal data can provide insight to many research questions and be used to emulate a clinical trial.[22]

With claims data, and observational data in general, the assignment of treatment is not random, and thus susceptible to confounding and selection bias. In practice, the clinician and patient decide among treatment options based on the patient's circumstances and overall health. Further, these same factors may also be associated with the health outcome of interest. If not properly incorporated, these factors will misrepresent and bias the true treatment effect comparison. While there are several approaches to handling confounding and selection bias available, propensity score-based methods are versatile in that they can be used for a variety of research questions and can be used for many different kinds of study designs and databases. Propensity score models can be particularly useful when there are many potential confounders and the outcome is rare. Propensity score approaches also prevent p-hacking of a desired result in the outcome model.[23] Thus, these methods have gained increasing popularity, especially for questions of comparative effectiveness in pharmacoepidemiologic and pharmacoeconomic research.

Lack of Reporting and Reproducibility

A downside to this rise in popularity is that the assumptions and critical steps for the propensity score-based methods are often ignored or unreported. This lack of reporting hinders other researchers' ability to replicate the findings. Ali et al[24] found 296 published medical papers in a 6-month period that reported use of a propensity score method. However, in their systematic review, they found that 194 (65.5%) did not report how variables were selected for the propensity model, and that only 177 (59.8%) reported test for balance of confounders between the two groups of comparison. Others have also noted common misuse of propensity methods.[25,26,27,28] Yao et al.[29] concluded in a recent systematic review of cancer studies that there is considerable room for improvement in reporting propensity analysis and offered guidelines for such reporting. Yet, some researchers are still not clear with their use of propensity methods and presentation in a scientific paper. For example, when comparing the effectiveness of allopurinol or febuxostat on reducing the risk of atrial fibrillation using Medicare data, Singh et al[30] matched subjects based on the propensity score. While they did report which variables were used for propensity construction and balance diagnostics after matching, many important details were not reported. Analysis questions arise, such as how the propensity score was calculated (logistic regression or otherwise), what distance measure was used to match subjects, if subjects were matched with or without replacement. These details are essential for researchers wishing to replicate the results reported. Additionally, even for many those researchers that did describe such methods, sensitively analysis to the results were often not reported. Propensity score methods do not account for unmeasured confounding, and sensitivity analyses can provide the reader with crucial information on the robustness of the findings. In many situations it is not clear what is the target estimand, for example, whether we are estimating the average treatment effect or the conditional average treatment effect.

Austin[31] provides a conceptual overview of propensity score methods from a foundational and introductory standpoint. Stuart et al[32] provide a general framework for using propensity methods with observational health care data, providing an example of effect estimation of drug monitoring programs for individuals with serious mental illness. Additionally, Brookhart et al[33] provide practical example when comparing the risk of

doi:10.20944/preprints202006.0295.v1

angioedema between two treatments for hypertension. While these papers offer an elegant and lucid exposition of the underlying principles, and are extremely important contribution to the literature, these overviews do not offer the reader complete practical guidance at every analysis step, as there remains a gap from methodological understanding to actual implementation. Further, these tutorials do not directly address the use of propensity methods for a range of outcomes commonly found in claims data, such as non-continuous, time to event or correlated outcomes. For example, a researcher may be interested in if a rare adverse event occurs or not (categorical) or monitoring a patient's disease progression over the course of several visits (correlated repeated measures and time to event outcomes). There are unique assumptions and considerations when using propensity methods for these different types of outcomes beyond those used for a simple continuous and normally distributed outcome. Additionally, there is need for a demonstrated sensitivity analysis after the treatment effect estimation to understand the strength of evidence supporting the results.

Therefore, there is need for a usable, simple and comprehensive tutorial for all stages of analysis when characterizing a binary treatment effect on various outcome types using claims data, with accompanying software code for each step. This paper outlines the use of three primary propensity score-based methods: Propensity Matching, Spline Adjustment, and Inverse Probability of Treatment Weighting (IPTW). The paper also details how to use each method to estimate average treatment effect for four common outcome types: 1) Binary, 2) Count, 3) Time to event, and 4) Longitudinally varying repeated measures. Finally, we conduct sensitivity analysis for two of the outcome types. To improve transparency for reproducibility and usage of the methods discussed, detailed R code is provided in an online version at https://rydaro.github.io/ .

To illustrate the entire process, we chose to study treatment patterns and treatment outcomes among patients with advanced stage prostate cancer from the Clinformatics TM Data Mart Database (OptumInsight, Eden Prairie, Minnesota). This database has a wealth of de-identified medical claims, pharmacy claims, inpatient confinement information, provider information, and socio-demographic information. Each outcome type is defined from emergency room visits (binary and count), time on treatment and in database (time to event), and prescription fills for opioids (repeated measures over time).

Causal Inference and Average Treatment Effect

Causal inference relies on the potential-outcomes framework, where each individual has a potential outcome under each possible treatment, with in reality only one actually assigned to him/her.[34] This framework allows researchers to clearly define what it means for an effect to be causal through the use of counterfactuals that can be viewed as missing data. Consider the case of two possible available treatments, a treatment of interest compared to another established treatment for the same disease, with a single outcome measured after treatment. We would define the causal effect of the treatment of interest as the difference between the actual and counterfactual outcomes in both treatment scenarios.

As described by Rubin,[34,35] many causal inference problems involve comparison of potential outcomes on the same (say i^{th}) individual. Define $Y_i(0)$ as the potential

doi:10.20944/preprints202006.0295.v1

outcome under the control treatment, and $Y_i(1)$ as the potential outcome under the active treatment of interest. We wish to know the treatment effect for each individual, typically defined as $Y_i(1) - Y_i(0)$, which cannot be estimated directly from the observed data because for each individual we observe either $Y_i(1)$ or $Y_i(0)$, but never both. If subject i actually received the active treatment, denoted by $T_i = 1,$ then $Y_i(1)$ is observed and $Y_i = Y_i(1)$; otherwise, $T_i = 0$, and we observe $Y_i = Y_i(0),$ under the stable unit treatment value and consistency assumptions. Often, researchers are interested in how patients receiving a specific treatment compares to a comparison group within a larger population. We can define the average treatment effect (ATE) as $E[Y_i(1) - Y_i(0)]$, which is the average treatment effect across the entire population.[36] In a randomized trial, we can estimate ATE as $E[Y_i(1) - Y_i(0)] = E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$ as randomization ensures that the treatment groups are balanced and hence $E[Y_i(a)] = E[Y_i(a)|T_i = a] = E[Y_i|T_i = a]$ for $a = 0{,}1$.[31,37] ATE can be defined on different scales, such as a ratio $\frac{E[Y_i|T_i = 1]}{E[Y_i|T_i = 0]}$ or odds ratio for binary outcomes $\frac{E[Y_i|T_i=1]/(1-E[Y_i|T_i=1])}{E[Y_i|T_i=0]/(1-E[Y_i|T_i=0])}$. We can also define the average treatment effect on the treated (ATT) as $E[Y_i(1) - Y_i(0)|T_i = 1]$ and the average treatment effect on the control (ATC) as $[Y_i(1) - Y_i(0)|T_i = 0]$ when a particular sub-population is of interest.

The standard method of estimating treatment effect for data from a randomized trial, or from observational data that is sufficiently balanced, is a general linear model with the treatment variable as the sole predictor:

$$g(\mu_i) = \beta_0 + \beta_1 T_i$$

where $\mu_i = E[Y_i|T_i]$ and β_1 is the parameter of interest for treatment comparison. In the simple linear regression case where $g()$ is the identity function, $\beta_1 = E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$. When using claims data, the mechanism behind treatment assignment is not random, and thus the treatment populations may differ greatly. Therefore $E[Y(1)|T = 1] \neq E[Y(1)]$ and $E[Y(0)|T = 0] \neq E[Y(0)]$ in general.[31] As a result, the estimate for β_1 will not equal the ATE because of confounding.

When confounders are present, a natural inclination would be to extend our outcome model to account for such confounders:

$$g(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}$$

However, β_1 in the multivariate adjustment model generally does not estimate ATE even if we have the correct confounders and the model is correctly specified, particularly when $g()$ is not a collapsible link function. One approach to estimate ATE is G-computation, which predicts the pair of potential outcomes for each individual.[38,39] The accompanying standard error can be computed using sandwich estimation.[40,41] While a valid analytical approach, it may be difficult for the researcher to specify the outcome model, as there may be limited understanding of the relationship between the outcome and each covariate. The notion of the propensity score, a unidimensional construct, offers an alternative analytical approach that may be more suitable. The researcher may have more subject matter knowledge to construct a proper propensity score model, may want to avoid unconscious bias of demonstrating a desired causal effect in the outcome models by choosing confounders to adjust for, or use the propensity score simply as a dimension reduction technique. Using the propensity score in analysis involves several steps that have to be exercised with care and caution, as outlined in Figure 1. Below, we briefly explain each analysis step, and demonstrate all steps in the prostate cancer treatment example.

Propensity Score Estimation

Proposed by Rosenbaum and Rubin,[42] the propensity score is defined as $e_i = Pr(T_i = 1|\boldsymbol{X}_i)$. The score can be interpreted as the probability a subject receives treatment, predicted from the confounding variables denoted as $\boldsymbol{X}_i$. Rosenbaum and Rubin[42] showed that conditional on the propensity score, an unbiased estimate of ATE can be obtained if the treatment is strongly ignorable. A treatment is strongly ignorable if two conditions are met: 1) $0 < P(T_i = 1|\boldsymbol{X}_i\ ) < 1$, 2) $(Y_i(0), Y_i(1)) \perp T_i|\boldsymbol{X}_i$.[42] The second of these assumptions is the “no unmeasured confounders” assumption. Thus, a critical assumption for use of the propensity score is that all variables that affect the outcome and treatment assignment are measured. If all confounding variables are identified and included, and the model is correctly specified, this score achieves covariate balance between treatment and control groups. More formally, the correct e_i satisfies that $T_i \perp \boldsymbol{X}_i|e_i$, removing the effect of the confounders from the treatment effect when we condition on e_i alone. We explain covariate balance in further detail in the balance assessment section. With the treatment groups more comparable, we can better characterize the treatment’s effect on the outcome of interest. We can estimate this probability using logistic regression, predicting treatment received from our observed covariates.

While logistic regression is commonly used to estimate this propensity score, researchers have expanded their attention beyond parametric models. Many have used machine learning methods such as boosted logistic regression, random forests, and neural networks.[43,44,45] Another method we highlight in this paper is the covariate balancing propensity score (CBPS) proposed by Imai and Ratkovic.[46]

Covariate Balancing Propensity Score (CBPS) is a generalized method of moments estimate that captures two characteristics of the propensity score, namely, as a covariate balancing score and as the conditional probability of treatment assignment.[46] This method is a more automated form of propensity score construction, in that it calculates the propensity score with the exact balancing goal in mind. Thus, CBPS provides a balancing score for each subject that ensures all covariates included in the CBPS construction are balanced. Therefore, CBPS is an efficient alternative to propensity score estimation by a parametric model. We do note that if using another estimation technique, the ultimate goal of the propensity model is not to predict treatment assignment, but to reduce bias by balancing covariates.[47]

Still, the treatment effect estimation methods are sensitive to misspecification of the propensity score model, and thus the variables and their functional forms used in this model can affect the estimation of average treatment effect. Many suggest including all variables at all associated with the outcome, while excluding those only associated with the treatment of interest, based on subject-matter knowledge.[33,48,49,50,51] Vanderweele[52] provides a comprehensive general guide to confounder selection in observational studies. The sensitivity analysis can show how estimates can change under many plausible propensity score models.

Using the Propensity Score

Once the propensity score is constructed, there are four basic ways to use the score in treatment effect estimation: 1) Stratification based on the propensity score, 2) Direct covariate adjustment using propensity score as a covariate in the outcome model, 3) Matching treatments and controls based on the propensity score (PM), and 4) Inverse probability treatment weighting on the propensity score (IPTW). Stratification ranks subjects by the estimated propensity score and splits them into mutually exclusive stratum (say, deciles). The treatment effect in each stratum (decile) can then be estimated and pooled to obtain an overall treatment effect.[53] We will not discuss stratification at length in the main paper as it is used less commonly,[54,55] and refer you to the online supplementary materials and website for further information regarding the implementation of this method. The rest of this paper will focus on the three routinely used methods: Spline Adjustment, Propensity Matching, and IPTW.

Spline Adjustment

The propensity score is the coarsest balancing score while the full list of confounders is the finest.[56] This approach is similar to the G-computation approach above, except the confounders in the outcome model are replaced with a single covariate of the predicted propensity score. The ATE is calculated from the predicted potential outcomes for each individual, and estimate the standard error using sandwich estimation.[38,39,40,41] Typically, the propensity score is fit with a smoothing function, such as a polynomial spline function,[56] allowing for a more flexible model that is also computationally fast and reliable.

Propensity Matching

The first method discussed is matching observations based on the propensity score to estimate ATT. Often, exactly identical scores do not exist across individuals, and thus matching requires a clear definition of "closeness" of propensity based on a measure of distance.[57,58] Stuart et al.[57] provide a comprehensive overview of the various matching methods available. In practice, it is common to do $1:1$ matching, where each individual in the treatment group is matched to a single individual in the comparison group, based on the predefined measure of closeness. This matching ratio can result in major loss of data, especially if the treatment groups are of very different sizes. An alternative is using $1:k$ matching, where k is a max number of controls. With a defined distance, called a caliper, all potential matches within the distance up to k will be matched. This allows for maximal efficiency of data while still reducing bias since all close matches are kept. There is little guidance on what caliper a researcher should specify; however, Austin[59] suggests a caliper of 0.2 standard deviations of the logit of the propensity score as a default choice that works well across scenarios. Matching typically estimates the ATT, though some packages and techniques can estimate ATE.[58]

Inverse Probability of Treatment Weighting (IPTW)

The next method we consider is the inverse probability of treatment (IPTW) proposed by Rosenbaum.[60] We can calculate the IPTW v_i as

$$v_i = \frac{T_i}{\hat{e}_i} + \frac{(1 - T_i)}{(1 - \hat{e}_i)}$$

where $\hat{e}_i$ is the estimated propensity score. These weights can be very unstable for extreme values of $\hat{e}_i$, so trimming (sometimes called truncating) these values away

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 June 2020 doi:10.20944/preprints202006.0295.v1

from the extreme is often practiced.[61,62] In this paper we assume values greater than .99 or less than .01 to be extreme, so these values are rounded to the less extreme boundaries. The construction of weights used here estimates ATE, and different constructions can be used for ATT and other effect estimates of interest.[62]

Balance Assessment

It is good practice to check if the chosen propensity method achieved its goal of balancing the covariates. While there are several balance diagnostics a common balance diagnostic originally proposed by Rosenbaum and Rubin[63] is the standardized difference (or standardized bias) for 1:1 matching, defined as

$$\frac{\overline{x}_t - \overline{x}_c}{s_p}$$

This is the difference in mean value of the covariate in the treatment group $\overline{x}_t$ vs. the control group $\overline{x}_c$, adjusting for variability s_p, where here we defined s_p as the pooled standard deviation of the two treatment groups, defined as $s_p = \sqrt{\frac{s_t^2+s_c^2}{2}}$.[64,65] This value is calculated for each covariate, with values closer to zero indicating better mean balance and potentially less bias. The measure can be calculated for both continuous and categorical indicator variables.[29,65] A lack of balance indicates that the propensity model may be incorrect, or that a different method should be used. There is no generally accepted threshold, although some suggest that the standardized difference should not be greater than 0.1.[64,65,66] We can modify this difference calculation for a different ration of matching, say $1{:}k$, using weights.[67,68] The weighted mean is defined as $\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i}$ and the weighted standard deviation is

$$s_w = \sqrt{\frac{\frac{\sum w_i(x_i - \overline{x}_w)^2}{\sum w_i}}{(\sum w_i)^2 - \sum {w_i}^2}}$$

where w_i is the weight for subject i. For 1:1 matching, all observations have equal weight. If $1{:}k$ matching is used, observations in the control treatment group have $1/k$ weights and treated observations have weights 1. For IPTW, the calculated weights can be used, so $v_i = w_i$ for each observation.[68] If sufficient balance is not achieved, the process of propensity score construction and balance assessment is repeated, by changing the functional form of the propensity model. An important note here is that a researcher can repeat this process until balance is achieved to a desired level. Experimenting with the model specification at this stage is preferable to post-hoc modification of the outcome model with ATE as a desired target, especially in terms of reproducibility of results.

Treatment Effect Estimation

Once sufficient balance has been achieved, one can estimate the average treatment effect using a general outcome model

$$g(\mu_i) = \beta_0 + \beta_1 T_i$$

This model can be used directly on the matched dataset if 1:1 matching is used. If $1{:}k$ matching or IPTW is used, the constructing weights need to be used as well. Weights can be incorporated in the same fashion as weights from a survey design, using robust standard error estimation to account for error in weight estimation.[61,68] For the spline adjustment model, ATE is estimated by G-computation (also called standardization) with direct variance calculation.[56] Once an estimate is obtained, it is often useful to run a sensitivity analysis to see how the estimate may change under different model specifications and understand how sensitive the result is to some unmeasured confounder.

Sensitivity Analysis

For the sensitivity analysis, we adapt the visualization tool of capturing vibration of effects from Patel et al.[69] to a universe of potential propensity score models. This visualization tool allows the researcher to see the results of many possible models, providing an overall understanding of the ATE estimate's robustness to model specifications with the observed set of confounders. To summarize sensitivity to an unobserved confounder, we calculate the estimate's E-value.[70] The E-value tells us the minimum value of the association parameter that an unobserved confounder must have with both the treatment and the outcome of interest. Put more simply, the E-value tells us how strong an unmeasured confounder must be to explain away a significant effect. A large E-value indicates that the significance of our estimate for ATE is robust to confounders not accounted for, whereas a small E-value is weak evidence of a significant causal relationship.

Example: Comparing Oral Hormone Therapy vs. Immunotherapy for Advanced Prostate Cancer

Many patients with advanced prostate cancer will receive a number of different therapies sequentially to try to control the disease and symptoms. The three different types of outcomes that we consider are based on what clinicians are typically interested in. Patients may have varying degrees of responsiveness and tolerance to different therapies during the period of treatment. For example, some patients who experience pain from their cancer will have pain relief after starting a treatment and thus require less opiates to manage their cancer. On the other hand, some patients will have poor tolerance of specific therapies and may experience exacerbation or development of comorbid conditions and seek emergency critical care. It is also important to note that a treatment is typically only continued for as long as it is effectively controlling the disease or symptom. Thus, the longer a patient is on a treatment, presumably the longer the duration of effective disease control on that treatment.

Cohort Definition and Data Preparation

We defined a cohort of men who received treatment for advanced prostate cancer, based on receiving one of four focus medications (abiraterone, enzalutamide, sipuleucel-T, docetaxel) known to have a survival benefit in men with advanced prostate cancer from January 2010 through June 2016 from the Clinformatics TM Data Mart Insurance Claims Database. The initial cohort included any patient over the age of 18 with a diagnosis of malignant neoplasm of the prostate, coded as “185” in ICD-9 and “C61” in ICD-10. We restricted our final cohort to include patients that were continuously enrolled in the plan for the 180 days before the first medication claim. Finally, we wished to compare first-line therapies between patients where first-line treatment was defined as the first medication given of the four focus medications. We then categorized patients given abiraterone or enzalutamide as a common oral therapy group. Thus, there are three final first-line treatment groups: 1) Immunotherapy, 2) Oral Therapy, and 3) Chemotherapy.

Define Treatment Effect (ATE)

Binary Outcome

We defined a binary outcome to be whether the patient had any emergency room (ER) visit within 60 days of the first pharmacy claim of the focus medications. ER visits were identified using both the provider and facility definition. The provider definition uses Current Procedural Technology (CPT) codes 99281-99285, and the facility definition uses revenue center codes 0450-0459, 098.[18,71]

Count Outcome

Using the previously defined ER visits, we counted the number of ER visits each patient had within 180 days from the first pharmacy claim as a count outcome. ATE is defined on the rate ratio scale

Time to Event Outcomes

We were also interested in the overall survival of patients; however, exact death dates were unavailable with this version of the data. We thus considered two other time to event outcomes as possible surrogates: time on treatment and time in database. Time on treatment was defined as the time from start of first medication to the last claim of any of the four focus medications, thus the event is stopping all focus treatment permanently. Time in database was defined as the time from start of the first medication to the last claim for that subject within the Clinformatics TM Data Mart Database for any medical-related issue. The last claim was identified by extracting the latest claim from each dataset, removing those after the enrollment end-date, and taking the maximum of those remaining. This definition of time in database could be considered a censored surrogate for death because we expect most patients to have medical needs until shortly before death. These two endpoints differ in that some individuals may have stopped treatment from a focus medication, yet still used medical services and managed pain beyond ending treatment, while others may have been treated continuously right up until death. Patients would be expected to have less total time on treatment if they had a highly resistant cancer that would not respond to any treatments (and thus treatments would not be continued if they were ineffective), or if they had severe toxicities to treatment that did not allow for continuation. Also, these endpoints differed across treatment groups, with those on oral therapy continuing treatment near the end of enrollment, whereas chemotherapy patients may stop a year or more before ending enrollment. ATE was defined as the mean difference in time, restricting to five years of follow-up.

Time Varying Repeated Measures Outcome

For the final longitudinal varying repeated measures outcome, we used opioid usage over time, calculated using prescription drug pharmacy claims. Common opioid drug types were identified and were converted into morphine milligram equivalents (MME) according to the Center for Disease Control conversion factors.[72] The total (MME) supply prescribed was calculated in 30-day periods, starting with the 30 days before the first-line of treatment, which was used as a baseline, and continuing at 30-day intervals for the duration of claims data available. Many patients with metastatic prostate cancer have pain from their disease that require opiates for pain control. Therefore, the level of MMEs may be a surrogate measure for disease burden, and disease response to treatment. ATE is defined as the mean difference in opioids prescribed at three specified time points: treatment start, 3 months after treatment start, and 6 months after treatment start.

Potential Confounder Selection

Potential confounders were identified using previous research explored factors associated with treatment and our outcomes of interest.[73,74] These include age, race, sociodemographic variables and comorbid conditions from Elixhauser Comorbidity Index and Clinical Classification Software,[75,76] all shown in Table 2. For more detail, see supplementary materials and website.

Propensity Analysis

Empirically identifying the optimal sequence of therapies through disease course is a complex problem due to sparse sample size. To determine which first-line treatment may lead to better outcomes regardless of which treatments a patient receives subsequently, we classified patients into one of the three categories of treatment that were prescribed first-line: oral therapy (abiraterone or enzalutamide), chemotherapy (docetaxel), or immunotherapy (sipuleucel-T). Since cabazitaxel and radium-223 were used infrequently as first-line treatments (n=110), we did not include patients who received cabazitaxel or radium-223 first-line in our analysis. We compared immunotherapy to oral therapy and compared immunotherapy to chemotherapy in two separate analyses. We chose immunotherapy as the reference group for both analyses, as it is the only treatment among the four included in the final analysis for which there is a clear treatment recommendation to be used in patients with minimally to asymptomatic metastatic castration-resistant prostate cancer. Our step-by-step example will primarily focus on the analysis process comparing immunotherapy to oral therapy and follow the same for comparing immunotherapy to chemotherapy.

Propensity Score Estimation

We can construct a model for treatment assignment, $T_i = 0$ if immunotherapy was given and $T_i = 1$ if oral therapy was given using logistic regression, and the CBPS method. We repeat the same analysis comparing immunotherapy to chemotherapy in a separate analysis. All potential confounders listed in the previous section were included. From the regression results, we can calculate the estimated propensity score for each subject e_i. The propensity score constructed from the CBPS approach was implemented through the R package *CBPS*.[46] The weights from this propensity score were used in the outcome models similar to the inverse probability weights. For chemotherapy estimation, the urologist variable was excluded as a confounder due to low cell counts.

Propensity Score Matching

To create a matched dataset, we used the R package *Matchit*.[77] We defined our distance with logistic regression using the “nearest neighbor” method select matches within a defined caliper distance of 0.2 standard deviations of the logit propensity score, with a variable matching ratio of $1:4$ within the defined caliper, without replacement. These matching specifications were chosen to ensure maximal efficiency of this data. By using variable matching, we allow multiple matches for a subject in the control group if several in the treatment group have close propensity scores by our defined distance measure. This allows us to retain more subjects in our analysis dataset than a standard 1:1 ration. The caliper was decided using an iterative process, where several calipers were assessed and the one providing the highest quality matched sample was kept, based on the standardized differences across the covariates.

Inverse Probability Treatment Weighting

Weights were created from both the logistic regression and CBPS estimated propensity scores using the formula described above. Some weights were unstable, so propensity scores greater that 0.99 were trimmed to 0.99, and scores below 0.01 were trimmed to 0.01. Trimmed weights were used for analysis.

Assessment of Covariate Balance

Each method can be assessed for successful reduction in standardized difference for the analysis sample. Figure 2 shows a plot of the standardized difference of the covariates between the immunotherapy group, and oral therapy group for CBPS, IPTW and propensity matching methods. We can see that the inverse weighted data and the matched sample reduced the standardized difference for many covariates, even if perfect balance was not achieved. Unsurprisingly, the CBPS weights have very low standardized differences in the means, as the weights are constructed to achieve this goal of exact matching. Here, we are assuming covariates have a linear relationship with the outcome, and thus checking means is sufficient. With balance among the covariates achieved, we can now begin treatment effect estimation.

Treatment Effect Estimation

Binary Outcome: Visit to the Emergency Room (ER) in 60 days

The first outcome of interest is whether a patient had an emergency room (ER) visit within the first 60 days of starting their treatment. Let $Y_i = 1$ if the i-th patient had an ER visit within the first 60 days of starting their first treatment, and $Y_i = 0$ if not. Thus, $\pi_i(1)$ is the probability an individual had an ER visit if they received oral therapy as first-line treatment, and $\pi_i(0)$ if they received the immunotherapy. We are interested in the odds ratio patient had an ER visit when treated with oral therapy to the odds a patient had an ER visit when treated with immunotherapy. We can model this odds ratio using a logistic regression model We cannot yet make any causal inferences from this model, as we haven't addressed the imbalance across the confounding variables. After running this model, we get an estimate of 0.75 (0.46,1.23), reported in Table 3. This odds ratio indicates that patients treated with oral therapy first line had 0.75 times the odds of an ER visit in 60 days than immunotherapy patients, before making adjustments.

Now we compare these results to our estimates of ATT and ATE. Since covariate balance is achieved, we can run the marginal logistic regression model on our propensity matched dataset, obtaining an estimate of 0.86 (0.51,1.45). Notice the larger confidence interval, as the matching process reduced the sample size. Next, we fit an outcome model on the full dataset that uses the propensity score directly as a covariate, using a spline function from the R package *splines*.78 ATE is calculate using the methods above, and we obtain an estimate of 0.83 (0.49, 1.41). Now, we can again fit the outcome model on the full dataset, now weighting each observation by the IPTW weights from the propensity scores estimated through logistic regression and the CBPS. Here, we use the same marginal model, using the weights for robust standard error estimation as described previously. We did so by using the R package *survey*.79 The estimates from these weighted models are 0.56 (0.26,1.23) and 0.55 (0.25 1.21). Finally, we report the multivariate adjustment model using G-computation with ATE of 0.80 (0.47, 1.37). None of these ORs were statistically significant, indicating that there may not be a significant difference in the odds of ER visits between these two treatment groups. When comparing immunotherapy and chemotherapy, the IPTW logistic regression and CBPS estimates are 1.51 (0.87, 2.61) and 1.85 (1.12,3.05), suggesting that chemotherapy patients may have a greater odds of an ER visit.

Count Outcome: Number of Emergency Room (ER) visits in 180 days

Next, we model our count outcome, the number of ER visits, where Y_i can take any positive integer values. We are interested in the rate ratio of the expected number of ER visits had all patients taken chemotherapy or oral therapy compared to immunotherapy. We can model that difference using a Poisson model with a log link. All models we fit in the binary outcome can be fit in a similar fashion to this count outcome, now considering the different link function and scale of ATE. Table 3 shows the results of each method for the count outcome. The models show that we can expect the same number of ER visits for patients who receive an oral therapy first-line vs. those who receive immunotherapy. For example, the matched ratio estimate is 1.00 (0.59,1.71), indicating the expected number of ER visits is the same for both treatment groups. However, we see a different pattern when comparing immunotherapy to chemotherapy, the matched ratio is 1.86 (1.15, 3.00), indicating that patients on chemotherapy have more ER visits.

Time to Event Outcomes: Time on Treatment and Time in Database

We will now discuss the time to events outcomes previously described. For each treatment group, we are interested in the difference in days from stopping all treatment, and the difference in days from total time in database. We can define these differences in terms of Restricted Mean Survival Time (RMST) within a given follow up window. We can calculate RMST, denoted μ_τ, as the area under the curve of the survival function:

$$\mu_\tau = \int_0^\tau S\,(t)dt$$

where $S(t)$ is the survival function, and τ is the parameter for restricted the follow-up time. We can then define our ATE estimate as $\mu_{\tau 1} - \mu_{\tau 0}$, or the difference in RMST between the treatment groups being compared. For example, when looking at our treatment time outcome comparing oral therapy to immunotherapy, we can interpret this ATE as “when restricting follow-up to τ days, patients given immunotherapy as first treatment will on average be treated μ_τ days longer than patients give oral therapy as first treatment.” For both of our outcomes, we estimated the survival function $S(t)$ using a Kaplan-Meier function, and choose $\tau = 1825$, restricting our follow-up time to 1,825 days - the equivalent of five years. We can estimate the difference in RMST using the package *survrm2*.[80] We can also obtain estimates of RMST with covariate adjustment[81] and with weights we calculate from the propensity score.[82]

Here, the matched estimate of -49 (-88, -9) shows that patients who receive an oral therapy first-line stopped treatment on average 49 days sooner than patients given the immunotherapy first-line, restricting to five years of follow-up. In other words, we'd expect patients who received an oral therapy first-line stop all treatment much sooner than patients who received immunotherapy as their first-line therapy. Again, looking at the matched estimates now comparing immunotherapy to chemotherapy, patients who received chemotherapy as first-line stopped all treatment an average of 167 (120, 214) days sooner than those patients who started on immunotherapy. In Table 3, we can see the time differences for ending enrollment as well, with matched estimates of -125 (-164, -96) comparing oral therapy to immunotherapy, and -177 (-224, -131) comparing chemotherapy to immunotherapy.

Longitudinally Varying Repeated Measures Outcome: Opioid Usage During Treatment

Lastly, in those patients who had an opioid prescribed at any time, we evaluated the longitudinally varying repeated measures outcome of opioids prescribed in MME per month for patients who had baseline opioid use before starting one of the focus treatments for their prostate cancer. Each patient included in this subset had baseline opioid prescriptions (30 days prior to start of treatment) as well as 180 days of opioids prescribed after initiation of treatment. The opioid prescriptions were defined in 30-day periods. We wish to model the trend and to test if there is any difference in mean opioid prescribing at any time point between treatment groups. We can model the quantity of opioids prescribed in MME Y_{ij} at the j^{th} 30-day period t_j for each individual i as:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 T_i + S(t_j) + S(t_j)T_i + \epsilon_{ij}$$

where $j = 1, .., n_i$, $n_i \in \{1,2,3,4,5,6,7\}$, $b_0 \sim N(0, \tau^2)$ and $\epsilon_i \sim MVN_{n_i}(0, \sigma^2 I_{n_i})$. Here, $S(t_j)$ is specified as a penalized regression spline with 3 degrees of freedom, allowing more flexible smooths for modeling the prescribing trend over time. We set the smooth in an interaction term to allow for different smooth trends for the immunotherapy and oral therapy treatment groups. Thus, the main parameter of interest tells us the difference in the mean opioid prescribing over time between the two groups. We can fit each of the methods in this outcome, adding covariates and smooths directly in the model, and fitting the model on a matched dataset. We use the R package *mgcv*.[83] Maindonald[84] also provides more detail on smooths when using GAM models. An important note when using IPTW and CBPS is that we are only weighting on the initial treatment, so at other time points the weights may bias the results. Also, we truncated the time to six months because many patients will only respond to or tolerate treatment for around six months before switching therapies to another focus treatment. Opiate use may parallel disease response to treatment in those who are started on opiates for their cancer. In other words, a patient's opiate use may decrease when their cancer is improving on treatment and subsequently increase when the cancer has become resistant to treatment. Pain management beyond six months from the initial treatment is unlikely related to that treatment as many patients have changed regimens or stopped treatment altogether. Any inferences using the full time period will be heavily biased by changing therapy or require advanced methods to handle switching treatments, such as marginal structure models.[85] Table 3 shows the estimated difference in mean opioid usage between groups at selected time points. For example, the difference in MME prescribed to an average individual in the immunotherapy group vs. the oral therapy group at treatment start is -83 MME (CI -391, 224) in the unadjusted model. In other words, among patients prescribed opioids, the average individual in the immunotherapy group treatment is predicted to have 83 more MME's of prescribed opioids than the average individual in the oral therapy group at treatment start; however, this difference is not significantly significant. This estimate changes 90 days post treatment start to -130 MME (CI -380, 121) demonstrating how the estimate varies across time. We did not detect any significant differences in opioid usage at any time point, for both the oral and immunotherapy comparison and the chemotherapy and immunotherapy comparison.

Sensitivity Analysis

We assessed robustness of the estimates by looking at the vibration of effects to many propensity score models based on observed set of confounders, and also calculating an E-value for unobserved confounding. For the binary outcome, we assessed the estimates to all possible propensity score models for three selected methods. Age was included as a baseline predictor in all models. E-values were calculated for the model that included the full covariate set. Figure 3 shows the results of these analyses.

Discussion

We have presented a very simple and standard use of propensity methods for estimating the causal effects of a treatment on the outcomes of interest that are routinely used. We showed methods that can make the comparison groups more balanced on a large number of characteristics, and thus provide more accurate estimates of possible causal relationships. To illustrate these methods, we analyzed treatment outcomes for different therapies used to treat patients with advanced prostate cancer. The results above showed that patients who received chemotherapy (docetaxel) first-line may have more frequent trips to the emergency room in the first six months compared to patients who receive immunotherapy as first-line therapy. The results also demonstrated that patients who received immunotherapy first-line may have longer total time on all treatments (first-line and subsequent treatments) than patients whose first-line therapy is an oral therapy or chemotherapy. Finally, among patients who already have a baseline opioid requirement for pain control when they initiated treatment for advanced prostate cancer, we saw higher average baseline requirements among those patients who were started on chemotherapy than those patients who were started on immunotherapy. However, patients in the chemotherapy group appeared to have better pain control after starting treatment than those patients started on immunotherapy.

There are inherent limitations to the data, as the Clinformatics TM Data Mart Database is designed for billing purposes and not for research. Thus, the data is subject to misclassification of diagnosis codes and is missing socioeconomic values for many individuals. Although we could not identify if an individual was correctly classified as having prostate cancer, we only included those that also had a pharmacy claim of one of the six focus medications which are primarily used for advanced prostate cancer. Those individuals with missing sociodemographic information were still included in the analysis and treated as a separate category.

A significant limitation to making any clinical conclusions about prostate cancer outcomes with the findings in this paper is that prostate cancer is a heterogeneous cancer, with a wide variation in prognosis and expected responses to therapy, even in the metastatic setting. Thus, a major unmeasured confounder when studying prostate cancer in claims data is the extent of disease at initiation of treatment. This unmeasured confounder may explain some of the observed effects on our outcomes. Claims can identify if a patient is metastatic but cannot identify the extent of their metastases. This limitation has significant implications if one were to clinically interpret the data. For example, when comparing opioid requirements and differences of opioid use among treatment groups, we cannot ascertain whether a patient is using opioids for their cancer or for another reason. It's possible that patients in the immunotherapy group who have a baseline opioid requirement may use opioids for a condition unrelated to their advanced prostate cancer, as opposed to patients in the chemotherapy or oral therapy group.

In addition, while we could identify when a patient visited the ER, we did not have the reason a patient visited the ER available. Patients may be presenting to the ER due to their disease, toxicities of the treatment, or another reason unrelated to their disease or treatment. These other un-related factors may be driving the large odds ratios observed between chemotherapy and immunotherapy patients. Identifying the fact that patients treated with chemotherapy first-line visit the ER more frequently may be signaling the fact that patients treated with chemotherapy first-line have more severe prostate cancer with more associated problems that require ER evaluation.

Some of these limitations are inherent to analyzing claims data. If we were able to control for disease severity at initiation of treatment, then an increased odds of visiting the ER would more reliably indicate a higher toxicity of therapy, or less control of disease from the treatment. Furthermore, since we cannot control for disease severity, we are not able to confidently say that patients who received immunotherapy are on treatment longer because of immunotherapy – we are only able to conclude that they remain on treatment longer. It's possible that patients started on immunotherapy have less aggressive disease at the start of therapy. However, interestingly, we did find that the increased time that patients in the immunotherapy group remained in database (potential surrogate for survival) compared to patients in the other two groups was longer than the differences we saw when comparing the amount of time on treatment. While impossible to conclude from these data, these data do suggest it's possible that patients who receive immunotherapy first-line may derive a longer-term benefit that is demonstrated even after all treatment is discontinued. For this comparison we used the dates from the last claim per individual as a censored time endpoint, as death records were unavailable. While the true death date is ideal, this endpoint is an underestimate for all prostate cancer patients and is a right censored measure of survival. These limitations are important for researchers to recognize, as the methods do provide conclusive interpretations when all confounders are controlled, however they do not overcome fundamental limitations of the data. Thus, researchers must be very cognizant of what variables are available, are used, and if they are adequate for causal interpretation.

There are also challenges and drawbacks to the methods used here. Propensity methods rely on correct specification of the propensity model. Here, we used a theoretical framework, pre-emptively specifying which variables are most associated with assignment of treatment, such as age, economic status, and pre-existing comorbid conditions. These variables were considered as potential confounders to both treatment and outcome assessment. Yet, we assessed many plausible propensity score models in our sensitivity analysis to assess the robustness of our findings. We were unable to account for all known confounders, and thus the propensity model may not have addressed all imbalance between groups. Our reporting of the E-value summarizes the sensitivity our results to unobserved confounding. Another potential limitation to this method is that we used a logistic regression model to calculate the propensity scores. While this model allows for natural interpretation of the variables included (which may still be of interest), it may be poor at predicting propensity in comparison to machine learning models.[43,44,45] Furthermore, the uncertainty around the propensity estimates is not accounted for in many outcome models, and thus lead to incorrect inference and confidence with the estimates.[32] Additionally, we effectively have three treatments of interest, yet we stratified the data to have two separate, independent analyses, of two treatment groups. This provided easier calculation and matching from propensity; however, segmenting may mis-specify the treatment allocation mechanisms, as in practice all options are available. Generalized propensity scores can be calculated for multiple categories, with the cost of considerably greater complexity.[86,87] Nonetheless, the methods are very useful for two clear treatment groups to be compared, and when there are many confounding variables.

Conclusion

In summary, the methods shown, and process outlined are very standard and routinely used tools for estimating causal effects from observed data in claims databases. It is important to note that these tools cannot perfectly answer causal questions, even with the most extensive data. There are assumptions that need to be met for causal interpretation of these estimates and they are often not verifiable from observed data. Careful consideration is required by the researchers as to what variables are confounding treatment and outcome, and what method and assumptions best fit the study. Adding sensitivity analysis to a study can add understanding to the robustness and generalizations of the results. We hope the extensive detail, documentation, and accompanying code aide researchers in their own studies and improve replication among these studies. The online implementable version of these steps for various types of outcomes with the accompanying tutorial guide is the most salient contribution of this paper.

References

1. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence — What Is It and What Can It Tell Us? *N Engl J Med* 2016; 375: 2293–2297.
2. Izurieta HS, Wu X, Lu Y, et al. Zostavax vaccine effectiveness among US elderly using real-world evidence: Addressing unmeasured confounders by using multiple imputation after linking beneficiary surveys with Medicare claims. *Pharmacoepidemiol Drug Saf* 2019; 28: 993–1001.
3. Noe MH, Shin DB, Doshi JA, et al. Prescribing Patterns Associated With Biologic Therapies for Psoriasis from a United States Medical Records Database. *J Drugs Dermatol* 2019; 18: 745–750.
4. Nidey N, Carnahan R, Carter KD, et al. Association of Mood and Anxiety Disorders and Opioid Prescription Patterns Among Postpartum Women. *Am J Addict*. Epub ahead of print 6 April 2020. DOI: 10.1111/ajad.13028.
5. O'Neal WT, Sandesara PB, Claxton JS, et al. Provider Specialty, Anticoagulation Prescription Patterns, and Stroke Risk in Atrial Fibrillation. *J Am Heart Assoc*; 7. Epub ahead of print 20 March 2018. DOI: 10.1161/JAHA.117.007943.
6. Desai RJ, Sarpatwari A, Dejene S, et al. Comparative effectiveness of generic and brand-name medication use: A database study of us health insurance claims. *PLoS Med*; 16. Epub ahead of print 1 March 2019. DOI: 10.1371/journal.pmed.1002763.
7. Jackevicius CA, Tu J V, Krumholz HM, et al. Comparative Effectiveness of Generic Atorvastatin and Lipitor® in Patients Hospitalized with an Acute Coronary Syndrome. *J Am Heart Assoc* 2016; 5: e003350.
8. Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets | FDA, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/best-practices-conducting-and-reporting-pharmacoepidemiologic-safety-studies-using-electronic (accessed 19 April 2020).
9. Motheral B, Brooks J, Clark MA, et al. A Checklist for Retrospective Database Studies—Report of the ISPOR Task Force on Retrospective Databases. *Value Heal* 2003; 6: 90–97.
10. Birnbaum HG, Cremieux PY, Greenberg PE, et al. Using Healthcare Claims Data for Outcomes Research and Pharmacoeconomic Analyses. *Pharmacoeconomics* 1999; 16: 1–8.
11. Johnson ML, Crown W, Martin BC, et al. Good Research Practices for Comparative Effectiveness Research: Analytic Methods to Improve Causal Inference from Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III. *Value Heal* 2009; 12: 1062–1073.
12. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf* 2017; 26: 1033–1039.
13. *Craig Dickstein and Renu Gehring Administrative Healthcare Data A Guide to Its Origin, Content, and Application Using SAS ®*.

14. *Framework for FDA's Real-World Evidence Program*, www.fda.gov (2018, accessed 25 April 2020).
15. Yu B, Kumbier K. Veridical data science. *Proc Natl Acad Sci* 2020; 117: 201901326.
16. Grimes DA. Epidemiologic Research Using Administrative Databases. *Obstet Gynecol* 2010; 116: 1018–1019.
17. Tyree PT, Lind BK, Lafferty WE. Challenges of using medical insurance claims data for utilization analysis. *Am J Med Qual* 2006; 21: 269–75.
18. Venkatesh AK, Mei H, E. Kocher K, et al. Identification of Emergency Department Visits in Medicare Administrative Claims: Approaches and Implications. *Acad Emerg Med* 2017; 24: 422–431.
19. Hoover KW, Tao G, Kent CK, et al. Epidemiologic Research Using Administrative Databases: Garbage In, Garbage Out. *Obstet Gynecol* 2011; 117: 729.
20. John Wilson B, President of Clinical Analytics V, Bock A. *Optum The benefit of using both claims data and electronic medical record data in health care analysis White Paper*, http://www.surescripts.com/pdfs/national-progress-report.pdf (accessed 11 March 2020).
21. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005; 58: 323–337.
22. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* 2016; 183: 758–764.
23. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: Analytic strategies using propensity scores. *Annals of Internal Medicine* 2002; 137: 693–695.
24. Ali MS, Groenwold RHH, Belitser S V., et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol* 2015; 68: 122–131.
25. Weitzen S, Lapane KL, Toledano AY, et al. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf* 2004; 13: 841–853.
26. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med* 2008; 27: 2037–2049.
27. D'ASCENZO F, CAVALLERO E, BIONDI-ZOCCAI G, et al. Use and Misuse of Multivariable Approaches in Interventional Cardiology Studies on Drug-Eluting Stents: A Systematic Review. *J Interv Cardiol* 2012; 25: 611–621.
28. Deb S, Austin PC, Tu J V., et al. A Review of Propensity-Score Methods and Their Use in Cardiovascular Research. *Can J Cardiol* 2016; 32: 259–265.
29. Yao XI, Wang X, Speicher PJ, et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. *JNCI J Natl Cancer Inst*; 109. Epub ahead of print 1 August 2017. DOI: 10.1093/jnci/djw323.
30. Singh JA, Cleveland JD. Comparative effectiveness of allopurinol and febuxostat for the risk of atrial fibrillation in the elderly: a propensity-matched analysis of Medicare claims data. *Eur Heart J*. Epub ahead of print 28 March 2019. DOI: 10.1093/eurheartj/ehz154.
31. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011; 46: 399–424.
32. Stuart EA, DuGoff E, Abrams M, et al. Estimating causal effects in observational studies

using Electronic Health Data: Challenges and (some) solutions. *EGEMS (Washington, DC)*; 1. Epub ahead of print 2013. DOI: 10.13063/2327-9214.1038.

33. Brookhart MA, Wyss R, Layton JB, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes* 2013; 6: 604–11.
34. Rubin DB. *ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES 1*, http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf (1974, accessed 8 January 2019).
35. Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Epub ahead of print 2005. DOI: 10.1198/016214504000001880.
36. Imbens GW. *Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review Author(s): Guido W. Imbens Source: The Review of Economics and Statistics NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY: A REVIEW**, https://faculty.smu.edu/millimet/classes/eco7377/papers/imbens 04.pdf (2004, accessed 8 January 2019).
37. Lunceford JK, Davidian M. *Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study*, https://www4.stat.ncsu.edu/~davidian/statinmed.pdf (accessed 8 January 2019).
38. Robins J. *A NEW APPROACH TO CAUSAL INFERENCE IN MORTALITY STUDIES WITH A SUSTAINED EXPOSURE PERIOD-APPLICATION TO CONTROL OF THE HEALTHY WORKER SURVIVOR EFFECT*.
39. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique | American Journal of Epidemiology | Oxford Academic, https://academic.oup.com/aje/article/173/7/731/104142 (accessed 25 April 2020).
40. Andersen BR. Introduction In : Modern Methods for Robust Regression. 2019; 1–6.
41. Susanti Y, Pratiwi H, Sulistijowati SH, et al. P A M ESTIMATION, S ESTIMATION, AND MM ESTIMATION IN ROBUST REGRESSION. *Int J Pure Appl Math* 2014; 91: 349–360.
42. Rosenbaum PR, Rubin DB. *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, http://www.stat.cmu.edu/~ryantibs/journalclub/rosenbaum_1983.pdf (1983, accessed 8 January 2019).
43. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2010; 29: 337–346.
44. Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* 2008; 17: 546–555.
45. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol* 2010; 63: 826–833.
46. Imai K, Ratkovic M. *Covariate balancing propensity score*. Epub ahead of print 2013. DOI: 10.1111/rssb.12027.
47. Wyss R, Ellis AR, Brookhart MA, et al. The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score. *Am J Epidemiol* 2014; 180: 645–655.

48. Rubin DB, Thomas N. *Matching Using Estimated Propensity Scores: Relating Theory to Practice*, https://www.jstor.org/stable/pdf/2533160.pdf?refreqid=excelsior%3A067c1e372596a3542acc597c244718a6 (1996, accessed 9 March 2019).
49. Rubin DB. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Ann Intern Med* 1997; 127: 757.
50. Perkins SM, Tu W, Underhill MG, et al. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf* 2000; 9: 93–101.
51. Wyss R, Girman CJ, LoCasale RJ, et al. Variable selection for propensity score models when estimating treatment effects on multiple outcomes: a simulation study. *Pharmacoepidemiol Drug Saf* 2013; 22: 77–85.
52. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*; 34. Epub ahead of print 2019. DOI: 10.1007/s10654-019-00494-6.
53. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *J Am Stat Assoc* 1984; 79: 516.
54. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* 2007; 26: 734–753.
55. Austin PC. The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. DOI: 10.1177/0272989X09341755.
56. Shi X, Wellman R, Heagerty PJ, et al. Safety surveillance and the estimation of risk in select populations: Flexible methods to control for confounding while targeting marginal comparisons via standardization. *Stat Med* 2020; 39: 369–386.
57. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics* 1985; 41: 103–16.
58. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci* 2010; 25: 1–21.
59. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat* 2011; 10: 150–161.
60. Rosenbaum PR. Model-Based Direct Adjustment. *J Am Stat Assoc* 1987; 82: 387–394.
61. Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. *PLoS One* 2011; 6: e18174.
62. Li F, Lock Morgan K, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting. *J Am Stat Assoc* ���� 2018; 113: 390–400.
63. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *Am Stat* 1985; 39: 33.
64. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009; 28: 3083–3107.
65. Normand S-LT, Landrum MB, Guadagnoli E, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *J Clin Epidemiol* 2001; 54: 387–398.

66. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. DOI: 10.1002/pds.1674.
67. Joffe MM, Ten Have TR, Feldman HI, et al. Model Selection, Confounder Control, and Marginal Structural Models. *Am Stat* 2004; 58: 272–279.
68. Morgan SL, Todd JJ. 6. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. *Sociol Methodol* 2008; 38: 231–282.
69. Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* 2015; 68: 1046–1058.
70. Van Der Weele TJ, Ding P. Sensitivity analysis in observational research: Introducing the E-Value. *Ann Intern Med* 2017; 167: 268–274.
71. Measure Methodology | CMS, https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology (accessed 25 April 2020).
72. Data Resources | Drug Overdose | CDC Injury Center, https://www.cdc.gov/drugoverdose/resources/data.html (accessed 25 April 2020).
73. Caram ME V., Wang S, Tsao P, et al. Patient and Provider Variables Associated with Systemic Treatment of Advanced Prostate Cancer. *Urol Pract* 2019; 6: 234–242.
74. Caram MEV, Ross R, Lin P, et al. Factors Associated With Use of Sipuleucel-T to Treat Patients With Advanced Prostate Cancer. *JAMA Netw open* 2019; 2: e192589.
75. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity Measures for Use with Administrative Data. *Med Care* 1998; 36: 8–27.
76. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version), https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp (accessed 25 April 2020).
77. Ho DE, Imai K, King G, et al. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Polit Anal* 2007; 15: 199–236.
78. splines package | R Documentation, https://www.rdocumentation.org/packages/splines/versions/3.6.2 (accessed 25 April 2020).
79. Lumley T, Maintainer`` M, Lumley'' T. *Package 'survey' Title Analysis of Complex Survey Samples*, http://r-survey.r-forge.r-project.org/survey/ (2019, accessed 22 August 2019).
80. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014; 32: 2380–2385.
81. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* 2014; 15: 222–233.
82. Conner SC, Sullivan LM, Benjamin EJ, et al. Adjusted restricted mean survival times in observational studies. *Stat Med* 2019; 38: 3832–3860.
83. *Package 'mgcv' Title Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, https://cran.r-project.org/web/packages/mgcv/mgcv.pdf (2018, accessed 9 January 2019).
84. Maindonald J. *Smoothing Terms in GAM Models*, http://wwwmaths.anu.edu.au/~johnm/r-book/xtras/lm-compute.pdf. (2010, accessed 8 January 2019).
85. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural

Models. *Am J Epidemiol* 2008; 168: 656–664.

86. Hirano K, Imbens GW. *The propensity score with continuous treatments*, http://www.math.mcgill.ca/dstephens/SISCR2017/Articles/HIrano-Imbens-2004.pdf (accessed 8 January 2019).

87. Austin PC. Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Stat Med* 2018; 37: 1874–1894.

Tables

Table 1. Outcome Characteristics

	Immunotherapy (N = 504)		Chemotherapy (N = 2,214)		Oral Therapy (N = 2,747)	
Binary Outcome	**Count**	**(%)**	**Count**	**(%)**	**Count**	**(%)**
ER Visit in 60 Days	22	(4.4)	182	(8.2)	100	(3.6)
Count Outcome	**Mean**	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**
ER Visits in 180 Days	0.13	(0.44)	0.23	(0.79)	0.12	(0.50)
Time to Event Outcome (days)	**Median**	**(Q1, Q3)**	**Median**	**(Q1, Q3)**	**Median**	**(Q1, Q3)**
Time on Treatment[1]	227	(29,638)	110	(43,338)	224	(83,462)
Time in Database[2]	414	(183,785)	256	(105,541)	291	(125,541)
Longitudinally Varying Repeated Measures Outcome	**Count**	**(%)**	**Count**	**(%)**	**Count**	**(%)**
Enrolled at 90 days	438	(87.0)	1707	(77.1)	2235	(81.4)
Enrolled at 180 days	381	(75.6)	1353	(61.1)	1788	(65.1)
Any Opioids Prescribed at Any Time	166	(32.9)	936	(42.3)	1281	(46.6)
Opioids at Baseline[3]	73	(14.5)	653	(29.5)	825	(30.0)
Opioids at 90 Days	87	(19.9)	427	(25.0)	578	(25.9)
Opioids at 180 Days	65	(17.1)	359	(26.5)	515	(28.8)
Patients Prescribed (morphine milligram equivalents, 30-day supply)	**Median**	**(Q1, Q3)**	**Median**	**(Q1, Q3)**	**Median**	**(Q1, Q3)**
Opioids at Treatment Start	112	(39,435)	241	(75,1052)	184	(72,674)
Opioids 90 Days from Treatment Start	87	(73,871)	427	(87,1182)	578	(83,887)
Opioids 180 Days from Treatment Start	391	(97,895)	406	(89,1448)	191	(60,667)

Table 1 Legend: Table 1 shows outcome characteristics across the three treatment groups: immunotherapy (sipuleucel-T), chemotherapy (docetaxel), and oral therapy (enzalutamide or abiraterone). ER is an abbreviation for emergency room. Q1 denotes first quartile of distribution, and Q3 denotes third quartile.

[1] Total time on treatment was defined as when the last of any focus treatment was recorded.

[2] Ending enrollment was defined as the last claim of any type.

[3] Opioids were identified from a list of generic brand names and converted into 30 day milligram morphine equivalents (MME) using the CDC compilation and conversion factors.

Table 2. Confounder Characteristics

	Immunotherapy (N = 504)		Chemotherapy (N = 2,214)		Oral Therapy (N = 2,747)	
Variable	**Count**	**(%)**	**Count**	**(%)**	**Count**	**(%)**
Age						
<55	14	(2.8)	93	(4.2)	62	(2.3)
55-64	87	(17.3)	329	(14.9)	341	(12.4)
65-74	194	(38.5)	915	(41.3)	769	(30.0)
≥75	209	(41.7)	876	(39.6)	1574	(57.3)
Race						
White	369	(73.2)	1,582	(71.5)	1,863	(67.8)
Asian	7	(1.4)	33	(1.5)	68	(2.5)
Black	62	(12.3)	284	(12.8)	376	(13.7)
Hispanic	22	(4.4)	127	(5.7)	252	(9.2)
Unknown	24	(8.8)	188	(8.5)	188	(6.8)
Education level						
No College	122	(24.2)	689	(31.1)	814	(29.6)
Some College or More	348	(69.0)	1400	(63.2)	1827	(66.5)
Unknown	34	(6.7)	124	(5.6)	105	(3.8)
Household income range						
<50k	148	(29.4)	798	(36.0)	997	(36.3)
50k-99k	164	(32.4)	656	(29.6)	862	(31.4)
>99k	119	(23.6)	431	(19.5)	527	(19.2)
Unknown	73	(14.5)	329	(14.6)	361	(13.1)
Geographic Region[1]						
New England	24	(4.8)	109	(5.0)	151	(5.5)
Middle Atlantic	37	(7.3)	134	(6.1)	257	(9.4)
South Atlantic	129	(25.6)	554	(25.0)	582	(21.2)
East North Central	76	(15.1)	305	(13.8)	403	(14.7)
East South Central	20	(4.0)	86	(3.9)	89	(3.2)
West North Central	63	(12.5)	386	(17.4)	137	(5.0)
West South Central	50	(9.9)	231	(10.4)	250	(9.1)
Mountain	75	(14.9)	221	(10.0)	302	(11.0)
Pacific	30	(6.0)	179	(8.1)	557	(20.3)
Unknown	0	(0.0)	9	(0.4)	19	(0.7)
Product						
HMO	128	(25.4)	797	(36.0)	991	(36.1)
PPO	36	(7.1)	181	(8.2)	208	(7.6)
Other	340	(67.5)	1,236	(55.9)	1,548	(56.4)
Metastatic						
Yes	474	(94.0)	2010	(90.8)	2,301	(83.8)
No	30	(6.0)	204	(9.2)	446	(16.2)
ASO						
Yes	96	(19.0)	344	(15.7)	434	(15.8)
No	408	(81.0)	1,866	(84.3)	2,313	(84.2)
Provider						
Urologist	167	(33.1)	4	(0.2)	318	(11.6)
Other/ Unknown	337	(66.9)	2209	(99.8)	2428	(88.4)
Comorbid Conditions						
Diabetes	154	(30.6)	593	(26.8)	802	(29.2)
Hypertension	362	(71.8)	1,479	(66.8)	1,920	(69.9)

Arrhythmia	86	(17.1)	398	(18.0)	640	(23.3)
CHF	42	(8.3)	180	(8.1)	359	(13.1)
Osteoporosis	55	(11.0)	114	(5.1)	204	(7.4)

Table 2 Legend

Characteristics of patients by first of focus treatment given: immunotherapy (sipuleucel-T), chemotherapy (docetaxel), and oral therapy (enzalutamide or abiraterone)
HMO, health maintenance organization; PPO, preferred provider organization; ASO, administrative services only (self-funded health plan); CHF, Congestive Heart Failure
[1]Geographic region:

- New England (NE): Connecticut (CT), Maine (ME), Massachusetts (MA), New Hampshire (NH), Rhode Island (RI), Vermont (VT)
- Middle Atlantic (MA): New Jersey (NJ), New York (NY), Pennsylvania (PA)
- East North Central (ENC): Illinois (IL), Indiana (IN), Michigan (MI), Ohio (OH), Wisconsin (WI)
- West North Central (WNC): Iowa (IA), Kansas (KS), Minnesota (MN), Missouri (MO), Nebraska (NE), North Dakota (ND), South Dakota (SD)
- South Atlantic (SA): Delaware (DE), Washington D.C. (DC), Florida (FL), Georgia (GA), Maryland (MD), North Carolina (NC), South Carolina (SC), Virginia (VA), West Virginia (WV)
- East South Central (ESC): Alabama (AL), Kentucky (KY), Mississippi (MS), Tennessee (TN)
- West South Central (WSC): Arkansas (AR), Louisiana (LA), Oklahoma (OK), and Texas (TX)
- Mountain (M): Arizona (AZ), Colorado (CO), Idaho (ID), Montana (MT), Nevada (NV), New Mexico (NM), Utah (UT), Wyoming (WY)

Pacific (PAC): Alaska (AK), California (CA), Hawaii (HI), Oregon (OR), Washington (WA)

Table 3: Estimates of Causal Treatment Effects Across Methods of Oral Therapies or Chemotherapy Compared to Reference Immunotherapy

	Non-Causal	**ATT**	**ATE**			
	Unadjusted Association	**Matched**	**Spline of Propensity Score**	**IPTW using Propensity Score from Logistic Regression**	**IPTW using Propensity Score from CBPS**	**Multivariate Adjustment**
Binary Outcome: Emergency Room visit in 60 days - Odds Ratio Scale						
Oral Therapy	0.75 (0.46,1.23)	0.89 (0.53,1.50)	0.83 (0.50, 1.38)	0.56 (0.26,1.23)	0.59 (0.28,1.22)	0.80 (0.47, 1.37)
Chemotherapy	**1.86** **(1.16, 2.97)**	**1.74** **(1.08, 2.80)**	**1.75** **(1.09,2.82)**	**1.79** **(1.09,2.93)**	**1.81** **(1.11,2.95)**	**1.70** **(1.03, 2.81)**
Count Outcome: Number of Emergency Room visits in 180 Days - Rate Ratio Scale						
Oral Therapy	0.92 (0.56,1.52)	1.00 (0.59,1.71)	0.99 (0.63,1.56)	0.87 (0.48,1.60)	0.88 (0.46,1.70)	0.96 (0.60, 1.53)
Chemotherapy	**1.87** **(1.36,2.58)**	**1.86** **(1.15 3.00)**	**1.72** **(1.13,2.61)**	**1.74** **(1.29, 2.57)**	**2.75** **(1.73, 4.38)**	**1.73** **(1.15, 2.58)**
Time to Event Outcome: Total Time on Treatment – Difference in Mean Days on Treatment from Immunotherapy (restricted to 5 years of follow-up)						
Oral Therapy	**-68** **(-106, -30)**	**-52** **(-92, -12)**	**-49** **(-88, -9)**	**-27** **(-45, -10)**	**-31** **(-48, -13)**	**-57*** **(-95, -19)**
Chemotherapy	**-135** **(-174, -96)**	**-164** **(-213, -117)**	**-167** **(-214, -120)**	**-164** **(-184, -144)**	**-139** **(-160, -119)**	**-135*** **(-174, -95)**
Time to Event Outcome: Total Time in Database - Difference in Mean Days in Database from Immunotherapy (restricted to 5 years of follow-up)						
Oral Therapy	**-146** **(-184, -109)**	**-130** **(-169, -90)**	**-125** **(-164, -96)**	**-107** **(-125, -90)**	**-116** **(-134, -98)**	**-124*** **(-226, -22)**
Chemotherapy	**-147** **(-186, -108)**	**-172** **(-200, -125)**	**-177** **(-224, -131)**	**-186** **(-207, -164)**	**-155** **(-176, -134)**	**-147*** **(-194, -101)**
Longitudinally Varying Repeated Measures Outcome: Opioids Prescribed in Morphine Milligram Equivalents per 30-day period (mg/30 days) for Patients Prescribed						
Difference in Mean mg/30 days, Oral Therapy from Immunotherapy						

Treatment Start	-83 (-391,224)	-144 (-464, 177)	-104 (-420, 212)	-211 (-846, 423)	-44 (-311,221)	-106 (-419,208)
90 Days	-130 (-380, 121)	-169 (-431, 94)	-151 (-412, 110)	-342 (-738,52)	14 (-220, 249)	-130 (-388, 128)
180 Days	-178 (-497, 141)	-263 (-599, 73)	-199 (-526, 128)	-469 (-1114,177)	-63 (-343,216)	-181 (-506, 144)
Difference in Mean mg/30 days Chemotherapy from Immunotherapy						
Treatment Start	187 (-155,530)	291 (-133, 716)	203 (-173, 578)	301 (-100, 702)	258 (-46, 563)	177 (191, 547)
90 Days	34 (-248,316)	97 (-252,447)	50 (-272, 373)	-64 (-415, 287)	44 (-229, 317)	25 (-290, 341)
180 Days	226 (-133, 586)	234 (-220,687)	242 (-150, 635)	112 (-298,521)	284 (-50, 619)	235 (-152, 622)

Table 3 Legend: Table or estimates and confidence intervals for the treatment effect on each outcome. Immunotherapy is the reference group for each treatment comparison. Estimates reported are unadjusted association (before any adjustments are used, so estimate is non-causal observed association), using a propensity matched dataset, adjusting for propensity score in the outcome model, inverse propensity score weighting (IPTW) and covariate balance propensity score (CBPS), and estimate from predicted outcomes use full covariate adjustment. For binary and count outcomes, multivariate adjustment estimates come from G-computation. For time to event outcome, multivariate estimates are difference in mean time, restricted to 5 years of follow-up time. For time-varying, estimates are difference in mean opioid morphine milligram equivalents at the designated time points.

*Adjustment covariates limited to age and race due to computational issues with full covariate set.