HHS Public Access

Author manuscript

Health Serv Outcomes Res Methodol. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Health Serv Outcomes Res Methodol. 2021 June ; 21(2): 206–228. doi:10.1007/s10742-020-00222-8.

Veridical Causal Inference using Propensity Score Methods for Comparative Effectiveness Research with Medical Claims

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]

[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

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 inference using these datasets require thoughtful modifications 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 using medical claims. This paper offers practical guidance to the analytical process, demonstrates methods for estimating causal treatment effects with propensity score methods for several types of outcomes common to such studies, such as binary, count, time to event and longitudinally-varying measures, and also aims to increase transparency and reproducibility of reporting of results from these investigations. 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.

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, rydaro@umich.edu.

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

Keywords

average treatment effect; covariate adjustment; insurance claims; hormone therapy; matching; prostate cancer; reproducibility; sensitivity analysis; veridical data science

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 (Sherman et al. 2016; Izurieta et al. 2019; Noe et al. 2019; Nidey et al. 2020; O'Neal et al. 2018). 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 medications 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 (Desai et al. 2019; Jackevicius et al. 2016). Although these large data sources offer a wealth of information, there are many challenges and drawbacks, such as measured and unmeasured confounding, selection bias, heterogeneity, missing values, duplicate records and misclassification of disease and exposures. 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 (FDA 2011; Motheral et al. 2003; Birnbaum et al. 1999; Johnson et al. 2009; Berger et al. 2017; Dickstein and Gehring, 2014; FDA 2018). We emulate newly introduced principles from the predictability, computability, and stability (PCS) framework for veridical data science (Yu et al. 2020) to examine comparative effectiveness research questions that administrative claims data can be used to address. We provide documentation and code in R Markdown file available online at https://rydaro.github.io/.

Healthcare claims data have been extensively criticized for its use in epidemiological research (Grimes 2010; Tyree et al. 2006). These types of data are wrinkled with issues such as outcome and covariate misclassification, missing data, and selection bias. For example, International Classification of Disease (ICD) codes are entered into administrative records by the care provider, often only for the purpose of billing, and thus certain diagnoses may be missed or overrepresented or may differ across providers (Tyree et al. 2006). There is no agreed upon algorithm for identifying widely used outcomes like Emergency Room visits, and thus many definitions across analysts and institutions are used (Venkatesh et al. 2017). While not as accurate as gold standard clinical trial data, these datasets are still valuable and sometimes the only source of real-world data for a wide variety of questions regarding drug utilization, effectiveness, and monitoring of adverse events (Hoover et al. 2011; Wilson and Bock 2012). 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 capture a more comprehensive picture of a patient's encounters with the healthcare system than standard electronic medical record (EHR) data alone (Schneeweiss et al. 2005), 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 with claims data can be addressed (FDA 2011; Motheral et al. 2003; Birnbaum et al. 1999; Johnson et al. 2009; Berger et al. 2017; Dickstein and Gehring, 2014; FDA 2018),and these large databases of longitudinal information can provide insight into many research questions and be used to complement/supplement/ emulate a clinical trial (Hernan and Robins, 2018). While there are several approaches to handling confounding bias are 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 approaches also prevent p-hacking of a desired result in the outcome model (Braitman and Rosenbaum 2002). Thus, these methods have gained increasing popularity, especially for questions of comparative effectiveness in pharmacoepidemiologic and pharmacoeconomic research. With counterfactual thinking and causal inference gaining popularity in the statistical and epidemiological literature, principled use of propensity score based methods in observational databases have become more common.

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. Many have noted common misuse or lack of reporting for propensity methods (Ali et al. 2015; Yao et al. 2017; Weitzen et al. 2004; Austin 2008; D'Ascenzo et al. 2012; Deb et al. 2016). Analysis questions arise, such as how the propensity score was calculated (logistic regression or otherwise), and even for many researchers who did describe such methods, sensitively analysis to the violation of assumptions or choice of the propensity score model 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.

Some have offered valuable tutorials on propensity score estimation (Garrido et al. 2014; Austin 2011; Stuart et al. 2013; Brookhart et al. 2013). 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 provide the reader a complete practical guidance for every analysis step, or a detailed sensitivity analysis framework to understand the strength of evidence supporting the results when model assumptions change. 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 annotated R software code for each step. This paper outlines the use of three primary propensity score-based methods: Spline Adjustment, Propensity Matching, and Inverse Probability of Treatment Weighting (IPTW) for comparing treatment effects with the goal of reducing bias due to confounding. 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/. 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.

Guideline for the Comparative Effectiveness Data Analysis Pipeline

Cohort Definition and Average Treatment Effect

The first stage of analysis, as shown in Figure 1, is cohort definition. The STROBE checklist for cohort studies provides guidelines for defining a cohort and research question for analysis (von Elm et al. 2007). Once a cohort is defined, comparative effectiveness research for that cohort relies on the potential-outcomes framework, which as described by Rubin (1975 and 2005), involves comparison of potential outcomes on the same (say i^{th}) individual for each treatment. Define $Y_i(0)$ as the potential 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. 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 (Imbens 2004). 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 $a = 0,1$ (Austin 2011b; Lunceford and Davidian 2017) . 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 comparison group (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 $g(x)$ 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 (Austin 2011b). 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 (Robins 1986; Snowden et al. 2011). The accompanying standard error can be computed using sandwich estimation (Andersen 2019; Susanti et al. 2014). 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.

Confounder Selection and Propensity Score Estimation

Proposed by Rosenbaum and Rubin (1983), 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 (1983) 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) The probability of treatment given the covariates is not exactly 0 or 1, and 2) each potential outcome is independent of treatment, conditional on the covariates. More formally, these two conditions are: $0 < P(T_i = 1|\boldsymbol{X}_i) < 1$, 2) $(Y_i(0), Y_i(1) \perp T_i|\boldsymbol{X}_i$ (Rosenbaum and Rubin 1983). 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. 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 (Lee et al. 2010; Setoguchi et al. 2008; Westreich 2010). Another method we highlight in this paper is the covariate balancing propensity score (CBPS) proposed by Imai and Ratkovic (2014).

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 (Imai and Ratkovic 2014). 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 (Wyss et al. 2014).

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 (Brookhart et al. 2006; Rubin and Thomas 1996; Perkins et al. 2000; Wyss et al. 2013). Vanderweele (2019) 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.

Application of the Constructed 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 groups to obtain an overall treatment effect (Rosenbaum and Rubin 1984). We will not discuss stratification at length in the main paper as it is used less commonly (Austin et al. 2007; Austin 2009b), but the online materials provide further information. 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 (Shi et al. 2020). 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 (Robins 1986; Snowden et al. 2011; Stefanski and Boos, 2002, Andersen 2019; Susanti et al. 2014). Typically, the propensity score is fit with a smoothing function, such as a polynomial spline function (Shi et al. 2020), allowing for a more flexible model that is also computationally fast and reliable.

Propensity Matching—Matching observations based on the propensity score to estimate ATT and is based on a measure of distance (Stuart et al. 2010; Rosenbaum and Rubin 1985a). Stuart et al. (2010) provide a comprehensive overview of the various matching methods available. In practice, it is common to do 1: k matching, where k is the specified 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 (2011a) 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 (Stuart et al. 2010).

Inverse Probability of Treatment Weighting (IPTW)—The next method we consider is the inverse probability of treatment weighting (IPTW) proposed by Rosenbaum (1987). We can calculate the weights 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 from the extreme is often practiced (Rosenbaum 1987; Lee et al. 2011). The construction of weights used here estimates ATE, and different constructions can be used for ATT and other effect estimates of interest (Lee et al. 2011).

Balance Assessment

It is good practice to check if the chosen propensity method achieved its goal of balancing the covariates. Although there are several balance diagnostics, a common balance diagnostic originally proposed by Rosenbaum and Rubin (1985b) is the standardized difference (or standardized bias) for 1:1 matching, defined as

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

This is the difference in mean value of the covariate in the treatment group $\bar{x}_t$ vs. the control group $\bar{x}_c$, adjusting for variability s_p, where here we defined s_p as the pooled standard deviation of the two treatment groups, defined as $s_p = \sqrt{\frac{s_t^2 + s_c^2}{2}}$ (Austin 2009a;Normand et al. 2001). 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 (Yao et al. 2017; Normand et al. 2001). 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 (Austin 2008b; Austin 2009a; Normand et al. 2001). In practice, researchers may also report variable descriptive statistics before and after matching. We can modify this difference calculation for a different ration of matching, say 1: *k,* using weights (Joffe et all. 2004; Morgan and Todd 2008; Austin 2008b). The weighted mean is defined as $\bar{x}_w = \frac{\Sigma\, w_i x_i}{\Sigma\, w_i}$ and the weighted standard deviation is

$$s_w = \sqrt{\frac{\frac{\Sigma\, w_i (x_i - \bar{x}_i)^2}{\Sigma\, w_i}}{(\Sigma\, w_i)^2 - \Sigma\, w_i 2}}$$

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 (Morgan and Todd 2008; Austin 2008b). 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. The 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 and Sensitivity Analysis

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 constructed 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 (Lee et al. 2001; Morgan and Todd 2008). For the spline adjustment model, ATE is estimated by G-computation with direct variance calculation via M-estimation (Stefanski and Boos, 2002). 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.

For the sensitivity analysis, we adapt a visualization tool of capturing vibration of effects from Patel et al. (2015) to a universe of potential propensity score models. This visualization tool allows the researcher to see the results of many possible models at once, providing an overall understanding of the treatment effect estimate's robustness to changing model specifications with the observed set of measured confounders. To summarize sensitivity to an unobserved/unmeasured confounder, we calculate the estimate's E-value (Van Der Weele and Ding 2017). The E-value captures the minimum value of the association parameter that an unobserved confounder must have with both the treatment and the outcome of interest to nullify the result regarding the treatment effect on the outcome. Put more simply, the E-value tells us how strong an unmeasured confounder must be to explain away a significant treatment effect.

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

Cohort Definition and Average Treatment Effect

The cohort was defined as men who received treatment for advanced prostate cancer at any time during January 2010 through June 2016, based on receiving one of four primary medications (abiraterone, enzalutamide, sipuleucel-T, docetaxel) known to have a survival benefit in men with advanced prostate cancer. Data were 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, and were continuously enrolled in the plan for at least 180 days before the first medication claim.

Treatments: We are interested in comparing first-line therapies. First-line treatment was defined as the first medication given of the four focus medications. We then categorized oral therapies as abiraterone or enzalutamide. Thus, there are three final first-line treatment groups: 1) Immunotherapy, 2) Oral Therapy, and 3) Chemotherapy. 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. The remainder of this example will only discuss the oral therapy comparison, as all methods are directly translatable, and we report all results for both analyses in the tables.

Binary and Count Outcomes: We defined the 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 the provider definition, Current Procedural Technology (CPT) codes 99281-99285, and the facility definition, which is revenue center codes 0450-0459, 098 (Vankatesh et al. 2017; CMS 2020a, CMS 2020b). ATE is defined on the odds ratio scale. 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 Outcome: Time on treatment, the time to event outcome, 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. ATE is defined in terms of Restricted Mean Survival Time (RMST) (Royston et al. 2013; Andersen 2010) within a five year 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 (five years). 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.

Longitudinally Varying Outcome: For the longitudinally varying 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 (CDC 2020). The average daily 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. 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. 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. 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. 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 (Cole and Hernan 2008).

Confounder Selection, Propensity Score Estimation and Balance Assessment

Confounders: Potential confounders were identified using previous research to identify factors associated with both treatment and outcomes (Hoffman et al. 2011; Ward et al. 2004; Caram et al. 2019a; Barocas and Penson 2010; Caram et al. 2019b). These include age, race, sociodemographic variables and comorbid conditions from Elixhauser Comorbidity Index and Clinical Classification Software (CMS 2020a, CMS 2020b), all shown in Table 2. From the table, we can see differences across treatments groups, especially age, geographic region, and provider type. These variables may inform treatment assignment and as such should be considered as potential confounders.

All potential confounders listed in Table 2 were included. For chemotherapy estimation, the urologist variable was excluded as a confounder due to low cell counts. Thus, the model treatment assignment is $T_i = 0$ if immunotherapy was given and $T_i = 1$ if oral therapy was given using both logistic regression and the CBPS method. Propensity score constructed from the CBPS approach was implemented through the R package *CBPS* (Imai and Ratkovic 2013). To create a matched dataset, we used the R package *Matchit* (Hoe et al. 2007). 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. Inverse weights were created, and propensity scores greater that 0.99 were trimmed to 0.99, and scores below 0.01 were trimmed to 0.01. 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. Here, we are assuming covariates have a linear relationship with the outcome, and thus checking means is sufficient. With covariate balance achieved to a desired level, we proceeded to treatment effect estimation.

Treatment Effect Estimation and Sensitivity Analysis

Now we compute our estimates of ATT and ATE and report results in Table 3. Because covariate balance is achieved, we can run the marginal logistic regression model directly on our propensity matched dataset. The spline model was implemented spline function from the R package *splines* (Bates et al. 2020). Models using the IPTW weights form the propensity

scores estimated from logistic regression and the CBPS used the R package *survey* (Lumley 2020). Finally, the ATE from multivariate adjustment model using G-computation is calculated with our own defined functions (shown in website).

Binary and Count Outcomes: For the binary outcome, we obtain an odds ratio of 0.83 (0.50,1.38) for oral therapy versus immunotherapy using the spline adjustment, and 0.56 (0.26,1.23) when using IPTW with the same propensity score. All models we fit for the binary outcome can be fit in a similar fashion to this count outcome, now considering the different link function and scale of ATE from above, also shown in Table 3. Here, we see a more consistent picture, with the spline adjustment rate ratio of 0.99 (0.63,1.56) and IPTW rate ratio of 0.87 (0.48,1.60).

Time to Event Outcome: For time on treatment, the difference in RMST was estimated using the package *survrm2* (Uno et al. 2014).[80] RMST was modified for covariate adjustment (Tian et al. 2014) and with weights calculated from the propensity score (Conner et al. 2019). In this case, the spline adjusted estimate for the difference in days on treatment within 5 years is −49 (−88,−9) days, suggesting immunotherapy patients stay on treatments longer than oral therapy. The IPTW estimate is −27 (−45, −10) days, now yielding a smaller interval than the spline adjusted for this outcome.

Longitudinally Varying Outcome: For the longitudinally varying opioid usage outcome we use the R package *mgcv* (Simon et al. 2018). The spline adjusted estimate for the difference in mean daily opioid prescribed 90 days after treatment start is −151 (−412, 110) MME. This suggests that on average, of those given opioids, patients starting with immunotherapy may be prescribed less opioids than those starting with oral therapy 90 days after treatment start. The confidence interval is noticeably wide. The IPTW estimate at this same time point is even less certain with −342 (−738,52).

Sensitivity Analysis

Across all outcomes, patients given Immunotherapy as the first line therapy have better outcomes on average, though not all estimates are significant. To assess the robustness of those that are “noteworthy” in terms of significance, sensitivity analysis is performed. A sample sensitivity analysis for the binary outcome is shown for three selected methods in Figure 3. Age was included as a baseline predictor in all models. E-values are reported for the model that included the full covariate set. The figure tells us that when comparing patients starting on oral therapy vs. immunotherapy, none of the propensity models considered resulted in a significant difference. However, the difference between chemotherapy and immunotherapy was significant. The E-values for these models are 3.05 for the spline adjustment and 3.17 for the IPTW, indicating a unobserved confounder that has this risk ratio with both this treatment assignment and ER visit outcome can explain away the significant result.

Discussion

We have presented propensity score methods for comparative effectiveness of a treatment on various types of health outcomes. We showed methods that can make the compared

treatment groups more balanced on a large number of characteristics, and thus provide more accurate estimates of possible causal relationships. There are inherent limitations to these data, as the Clinformatics TM Data Mart Database is generated from information collected 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 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. This method comes with assumptions about the missing data mechanism that, if violated, could bias ATE. There are more advanced methods for missing data analysis for a propensity score analysis that others discuss more extensively (D'Agostino et al. 2001; D'Agostino and Rubin, 2000).

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. 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 from this data, and thus the propensity model may not have addressed all imbalance between groups. Our reporting of the E-value summarizes the sensitivity of our results to unobserved confounding. Another potential limitation to 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 (Lee et al. 2010; Setoguchi et al. 2008; Westreich 2010). Furthermore, the uncertainty around the propensity estimates is not accounted for in many outcome models, especially when using propensity score matching, and thus lead to incorrect inference and confidence with the estimates (Stuart et al. 2013). 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 (Hirano and Imbens 2005; Austin 2018). Nonetheless, the methods are very useful for two clear treatment groups to be compared, and when there are many confounding variables. There are also computational challenges when using R for these complex analyses on a large dataset, but the flexibility of custom code and available packages far outweigh this cost. Finally, our estimates for ATE varied across the methods demonstrated, and it is impossible to know which method performed best for each outcome when the true ATE and true propensity model are unknown. Our goal was to showcase the available options while recognizing each method comes with its own limitations. We recommend reporting the sensitivity analysis for transparency with researcher decision error surrounding ATE estimation.

Conclusion

In summary, the methods shown here outline a standard process for conducting comparative effectiveness research in claims databases. It is important to note that these tools cannot perfectly answer comparative effectiveness questions, even with the most extensive 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement of Research Support:

Dr. Mukherjee is funded by the University of Michigan Comprehensive Cancer Center Core Grant - P30 CA 046592, University of Michigan Precision Health Initiative and NSF DMS grant 1712933.

Dr. Caram is funded by a Prostate Cancer Foundation Young Investigator Award

References

Ali M. Sanni, Groenwold Rolf H.H., Belitser Svetlana V., Pestman Wiebe R., Hoes Arno W., Roes Kit C.B., de Boer Anthonius, and Klungel Olaf H.. 2015. “Reporting of Covariate Selection and Balance Assessment in Propensity Score Analysis Is Suboptimal: A Systematic Review.” Journal of Clinical Epidemiology 68 (2): 122–31. 10.1016/J.JCLINEPI.2014.08.011. [PubMed: 25579639]

Andersen Robert.: Modern Methods for Robust Regression 1–6. (2019)

Andersen PK, Perme MP. Pseudo-observations in survival analysis. Stat Methods Med Res. 2010;19(1):71–99. doi:10.1177/0962280209105020 [PubMed: 19654170]

Austin Peter C. 2008a. “A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003.” Statistics in Medicine 27 (12): 2037–49. 10.1002/sim.3150. [PubMed: 18038446]

Austin Peter C. 2009a. “Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples.” Statistics in Medicine 28 (25): 3083–3107. 10.1002/sim.3697. [PubMed: 19757444]

Austin Peter C. 2009b. “The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies.” Edited by McDonald Kathryn M.. Medical Decision Making 29 (6): 661–77. 10.1177/0272989X09341755. [PubMed: 19684288]

Austin Peter C. 2011a. “Optimal Caliper Widths for Propensity-score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies.” Pharmaceutical Statistics 10 (2): 150–61. 10.1002/PST.433. [PubMed: 20925139]

Austin Peter C., Grootendorst Paul, and Anderson Geoffrey M.. 2007. “A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study.” Statistics in Medicine 26 (4): 734–53. 10.1002/sim.2580. [PubMed: 16708349]

Austin PC (2008b), Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidem. Drug Safe, 17: 1218–1225. doi:10.1002/pds.1674

Austin Peter C. 2011b. “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.” Multivariate Behavioral Research 46 (3): 399–424. 10.1080/00273171.2011.568786. [PubMed: 21818162]

Austin Peter C. 2018. “Assessing the Performance of the Generalized Propensity Score for Estimating the Effect of Quantitative or Continuous Exposures on Binary Outcomes.” Statistics in Medicine 37 (11): 1874–94. 10.1002/sim.7615. [PubMed: 29508424]

Barocas DA and Penson DF (2010), Racial variation in the pattern and quality of care for prostate cancer in the USA: mind the gap. BJU International, 106: 322–328. doi:10.1111/j.1464-410X.2010.09467.x [PubMed: 20553251]

Bates Douglas, and Venables William. 2020 “Splines Package | R Documentation.” Accessed April 25, 2020. https://www.rdocumentation.org/packages/splines/versions/3.6.2.

Berger Marc L., Sox Harold, Willke Richard J., Brixner Diana L., Eichler Hans-Georg, Goettsch Wim, Madigan David, et al. 2017. “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.” Pharmacoepidemiology and Drug Safety 26 (9): 1033–39. 10.1002/pds.4297. [PubMed: 28913966]

Birnbaum Howard G., Cremieux Pierre Y., Greenberg Paul E., Jacques LeLorier JoAnn Ostrander, and Venditti Laura. 1999. “Using Healthcare Claims Data for Outcomes Research and Pharmacoeconomic Analyses.” PharmacoEconomics 16 (1): 1–8. 10.2165/00019053-199916010-00001.

Braitman Leonard E., and Rosenbaum Paul R.. 2002. “Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores.” Annals of Internal Medicine. American College of Physicians. 10.7326/0003-4819-137-8-200210150-00015.

Brookhart M. Alan, Schneeweiss Sebastian, Rothman Kenneth J., Glynn Robert J., Avorn Jerry, and Starmer Til. 2006. “Variable Selection for Propensity Score Models.” American Journal of Epidemiology 163 (12): 1149–56. 10.1093/aje/kwj149. [PubMed: 16624967]

Brookhart M Alan, Wyss Richard, Layton J Bradley, and Starmer Til. 2013. “Propensity Score Methods for Confounding Control in Nonexperimental Research.” Circulation. Cardiovascular Quality and Outcomes 6 (5): 604–11. 10.1161/CIRCOUTCOMES.113.000359. [PubMed: 24021692]

Caram Megan E. V., Wang Shikun, Tsao Phoebe, Griggs Jennifer J., Miller David C., Hollenbeck Brent K., Lin Paul, and Mukherjee Bhramar. 2019a. “Patient and Provider Variables Associated with Systemic Treatment of Advanced Prostate Cancer.” Urology Practice 6 (4): 234–42. 10.1097/UPJ.0000000000000020. [PubMed: 31276025]

Caram Megan E.V., Ross Ryan, Lin Paul, and Mukherjee Bhramar. 2019b. “Factors Associated with Use of Sipuleucel-T to Treat Patients With Advanced Prostate Cancer.” JAMA Network Open 2 (4): e192589. 10.1001/jamanetworkopen.2019.2589. [PubMed: 31002323]

CDC. n.d. “Data Resources | Drug Overdose.” Accessed April 25, 2020a. https://www.cdc.gov/drugoverdose/resources/data.html.

CMS. n.d. “Measure Methodology.” Accessed April 25, 2020b. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.

Cole SR, and Hernan MA. 2008. “Constructing Inverse Probability Weights for Marginal Structural Models.” American Journal of Epidemiology 168 (6): 656–64. 10.1093/aje/kwn164. [PubMed: 18682488]

Conner Sarah C., Sullivan Lisa M., Benjamin Emelia J., LaValley Michael P., Galea Sandro, and Trinquart Ludovic. 2019. “Adjusted Restricted Mean Survival Times in Observational Studies.” Statistics in Medicine 38 (20): 3832–60. 10.1002/sim.8206. [PubMed: 31119770]

D’Agostino R, Lang W, Walkup M et al. Examining the Impact of Missing Data on Propensity Score Estimation in Determining the Effectiveness of Self-Monitoring of Blood Glucose (SMBG). Health Services & Outcomes Research Methodology 2, 291–315 (2001). 10.1023/A:1020375413191

D’Agostino RB Jr & Donald Rubin. (2000). Estimating and Using Propensity Scores with Partially Missing Data. Journal of the American Statistical Association. 95. 749–759. doi: 10.2307/2669455.

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(6):611–621. doi:10.1111/j.1540-8183.2012.00753.x [PubMed: 22882654]

Deb Saswata, Austin Peter C., Tu Jack V., Ko Dennis T., Mazer C. David, Kiss Alex, and Fremes Stephen E.. 2016. "A Review of Propensity-Score Methods and Their Use in Cardiovascular Research." Canadian Journal of Cardiology 32 (2): 259–65. 10.1016/J.CJCA.2015.05.015.

Desai Rishi J., Sarpatwari Ameet, Dejene Sara, Khan Nazleen F., Lii Joyce, Rogers James R., Dutcher Sarah K., et al. 2019. "Comparative Effectiveness of Generic and Brand-Name Medication Use: A Database Study of Us Health Insurance Claims." PLoS Medicine 16 (3). 10.1371/journal.pmed.1002763.

Dickstein Craig, and Gehring Renu. 2014. Administrative Healthcare Data A Guide to Its Origin, Content, and Application Using SAS ®. SAS Institute.

Elixhauser Anne, Steiner Claudia, Harris D. Robert, and Coffey Rosanna M.. 1998. "Comorbidity Measures for Use with Administrative Data." Medical Care 36 (1): 8–27. 10.1097/00005650-199801000-00004. [PubMed: 9431328]

FDA. 2011. "Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets."

FDA. 2018. "Framework for FDA's Real-World Evidence Program."

Garrido MM, Kelley AS, Paris J, Roza K, Meier DE, Morrison RS and Aldridge MD (2014), Methods for Constructing and Assessing Propensity Scores. Health Serv Res, 49: 1701–1720. 10.1111/1475-6773.12182 [PubMed: 24779867]

Grimes David A. 2010. "Epidemiologic Research Using Administrative Databases." Obstetrics & Gynecology 116 (5): 1018–19. 10.1097/AOG.0b013e3181f98300. [PubMed: 20966682]

HCUP. n.d. "Clinical Classifications Software (CCS) for ICD-10-PCS (Beta Version)." Accessed April 25, 2020. https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.

Hernán Miguel A., and Robins James M.. 2016. "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available." American Journal of Epidemiology 183 (8): 758–64. 10.1093/aje/kwv254. [PubMed: 26994063]

Hirano Keisuke, and Imbens Guido W.. 2005. "The Propensity Score with Continuous Treatments." In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family, 73–84. Wiley Blackwell. 10.1002/0470090456.ch7.

Ho Daniel, Imai Kosuke, King Gary, and Stuart Elizabeth. 2018. "Package 'MatchIt' Title Nonparametric Preprocessing for Parametric Causal Inference." 10.1093/pan/mpl013.

Hoffman RM, Gilliland FD, Eley JW, et al. Racial and ethnic differences in advanced-stage prostate cancer: the Prostate Cancer Outcomes Study. J Natl Cancer Inst. 2001;93(5):388–395. 10.1093/jnci/93.5.388 [PubMed: 11238701]

Hoover Karen W., Tao Guoyu, Kent Charlotte K., and Aral Sevgi O.. 2011. "Epidemiologic Research Using Administrative Databases: Garbage In, Garbage Out." Obstetrics & Gynecology 117 (3): 729. 10.1097/AOG.0b013e31820cd18a. [PubMed: 21343778]

Imai K and Ratkovic M (2014), Covariate balancing propensity score. J. R. Stat. Soc. B, 76: 243–263. 10.1111/rssb.12027

Imbens Guido. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. The Review of Economics and Statistics. 86. 4–29. 10.1162/003465304323023651.

Izurieta Hector S., Wu Xiyuan, Lu Yun, Chillarige Yoganand, Wernecke Michael, Lindaas Arnstein, Pratt Douglas, et al. 2019. "Zostavax Vaccine Effectiveness among US Elderly Using Real-World Evidence: Addressing Unmeasured Confounders by Using Multiple Imputation after Linking Beneficiary Surveys with Medicare Claims." Pharmacoepidemiology and Drug Safety 28 (7): 993–1001. 10.1002/pds.4801. [PubMed: 31168897]

Jackevicius Cynthia A, Tu Jack V, Krumholz Harlan M, Austin Peter C, Ross Joseph S, Stukel Therese A, Koh Maria, Chong Alice, and Ko Dennis T. 2016. "Comparative Effectiveness of Generic Atorvastatin and Lipitor® in Patients Hospitalized with an Acute Coronary Syndrome." Journal of the American Heart Association 5 (4): e003350. 10.1161/JAHA.116.003350. [PubMed: 27098970]

Joffe Marshall M, Ten Have Thomas R, Feldman Harold I, and Kimmel Stephen E. 2004. “Model Selection, Confounder Control, and Marginal Structural Models.” The American Statistician 58 (4): 272–79. 10.1198/000313004X5824.

Johnson Michael L., Crown William, Martin Bradley C., Dormuth Colin R., and Siebert Uwe. 2009. “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 in Health 12 (8): 1062–73. 10.1111/J.1524-4733.2009.00602.X. [PubMed: 19793071]

Lee Brian K., Lessler Justin, and Stuart Elizabeth A.. 2011. “Weight Trimming and Propensity Score Weighting.” Edited by Biondi-Zoccai Giuseppe. PLoS ONE 6 (3): e18174. 10.1371/journal.pone.0018174. [PubMed: 21483818]

Lee Brian K, Lessler Justin, and Stuart Elizabeth A. 2010. “Improving Propensity Score Weighting Using Machine Learning.” Statistics in Medicine 29 (3): 337–46. 10.1002/sim.3782. [PubMed: 19960510]

Li Fan, Morgan Kari Lock, and Zaslavsky Alan M.. 2018. “Balancing Covariates via Propensity Score Weighting.” Journal of the American Statistical Association 113 (521): 390–400. 10.1080/01621459.2016.1260466.

Lumley Thomas. 2020. R package “Survey” https://cran.r-project.org/web/packages/survey/index.html

Lunceford Jared K, and Davidian Marie. 2017. “Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study.”

Maindonald J: Smoothing terms in GAM models (2010)

Morgan Stephen L., and Todd Jennifer J.. 2008. “6. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects.” Sociological Methodology 38 (1): 231–82. 10.1111/j.1467-9531.2008.00204.x.

Motheral Brenda, Brooks John, Clark Mary Ann, Crown William H., Davey Peter, Hutchins Dave, Martin Bradley C., and Stang Paul. 2003. “A Checklist for Retrospective Database Studies—Report of the ISPOR Task Force on Retrospective Databases.” Value in Health 6 (2): 90–97. 10.1046/J.1524-4733.2003.00242.X. [PubMed: 12641858]

Nidey Nichole, Carnahan Ryan, Carter Knute D., Strathearn Lane, Bao Wei, Greiner Andrea, Jelliffee-Pawlowski Laura, Tabb Karen M., and Ryckman Kelli. 2020. “Association of Mood and Anxiety Disorders and Opioid Prescription Patterns Among Postpartum Women.” The American Journal on Addictions, 4. 10.1111/ajad.13028.

Noe Megan H., Shin Daniel B., Doshi Jalpa A., Margolis David J., and Gelfand Joel M.. 2019. “Prescribing Patterns Associated With Biologic Therapies for Psoriasis from a United States Medical Records Database.” Journal of Drugs in Dermatology : JDD 18 (8): 745–50. [PubMed: 31424706]

Normand Sharon-Lise T., Landrum Mary Beth, Guadagnoli Edward, Ayanian John Z., Ryan Thomas J., Cleary Paul D., and McNeil Barbara J.. 2001. “Validating Recommendations for Coronary Angiography Following Acute Myocardial Infarction in the Elderly: A Matched Analysis Using Propensity Scores.” Journal of Clinical Epidemiology 54 (4): 387–98. 10.1016/S0895-4356(00)00321-8. [PubMed: 11297888]

O’Neal Wesley T., Sandesara Pratik B., Claxton J Neka S., MacLehose Richard F., Chen Lin Y., Bengtson Lindsay G. S., Chamberlain Alanna M., Norby Faye L., Lutsey Pamela L., and Alonso Alvaro. 2018. “Provider Specialty, Anticoagulation Prescription Patterns, and Stroke Risk in Atrial Fibrillation.” Journal of the American Heart Association 7 (6). 10.1161/JAHA.117.007943.

Patel Chirag J., Burford Belinda, and Ioannidis John P.A.. 2015. “Assessment of Vibration of Effects Due to Model Specification Can Demonstrate the Instability of Observational Associations.” Journal of Clinical Epidemiology 68 (9): 1046–58. 10.1016/j.jclinepi.2015.05.029. [PubMed: 26279400]

Perkins Susan M., Tu Wanzhu, Underhill Michael G., Zhou Xiao-Hua, and Murray Michael D.. 2000. “The Use of Propensity Scores in Pharmacoepidemiologic Research.” Pharmacoepidemiology and Drug Safety 9 (2): 93–101. 10.1002/(SICI)1099-1557(200003/04)9:2<93::AID-PDS474>3.0.C0;2-I. [PubMed: 19025807]

Robins James. 1986. “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period-Application to Control of the Healthy Worker Survivor Effect.” Mathematical Modelling 7 (9–12): 1393–1512. 10.1016/0270-0255(86)90088-6.

Rosenbaum PR, and Rubin DB. 1985a. “The Bias Due to Incomplete Matching.” Biometrics 41 (1): 103–16. 10.2307/2530647. [PubMed: 4005368]

Rosenbaum Paul R. 1987. “Model-Based Direct Adjustment.” Journal of the American Statistical Association 82 (398): 387–94. 10.1080/01621459.1987.10478441.

Rosenbaum Paul R., and Rubin Donald B.. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” Journal of the American Statistical Association 79 (387): 516. 10.2307/2288398.

Rosenbaum Paul R. 1985b. “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.” The American Statistician 39 (1): 33. 10.2307/2683903.

Rosenbaum Paul R, and Rubin Donald B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” Biometrika. Vol. 70.

Royston P, Parmar MK Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 13, 152 (2013). 10.1186/1471-2288-13-152 [PubMed: 24314264]

Rubin Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized And Nonrandomized Studies 1.” Journal of Educational Psychology. Vol. 66.

Rubin Donald B. 2005. “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions.” 10.1198/016214504000001880.

Rubin Donald B, and Thomas Neal. 1996. “Matching Using Estimated Propensity Scores: Relating Theory to Practice.” Vol. 52.

Schneeweiss Sebastian, and Avorn Jerry. 2005. “A Review of Uses of Health Care Utilization Databases for Epidemiologic Research on Therapeutics.” Journal of Clinical Epidemiology 58 (4): 323–37. 10.1016/j.jclinepi.2004.10.012. [PubMed: 15862718]

Setoguchi Soko, Schneeweiss Sebastian, Brookhart M. Alan, Glynn Robert J., and Cook E. Francis. 2008. “Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study.” Pharmacoepidemiology and Drug Safety 17 (6): 546–55. 10.1002/pds.1555. [PubMed: 18311848]

Sherman Rachel E., Anderson Steven A., Dal Pan Gerald J., Gray Gerry W., Gross Thomas, Hunter Nina L., LaVange Lisa, et al. 2016. “Real-World Evidence — What Is It and What Can It Tell Us?” New England Journal of Medicine 375 (23): 2293–97. 10.1056/NEJMsb1609216.

Shi Xu, Wellman Robert, Heagerty Patrick J., Nelson Jennifer C., and Cook Andrea J.. 2020. “Safety Surveillance and the Estimation of Risk in Select Populations: Flexible Methods to Control for Confounding While Targeting Marginal Comparisons via Standardization.” Statistics in Medicine 39 (4): 369–86. 10.1002/sim.8410. [PubMed: 31823406]

Snowden Jonathan, Rose Sherri, and Mortimer Kathleen. 2011. “Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique.” American Journal of Epidemiology. 2011. https://academic.oup.com/aje/article/173/7/731/104142.

Stefanski Leonard A., and Boos Dennis D.. “The Calculus of M-Estimation.” The American Statistician, vol. 56, no. 1,2002, pp. 29–38. JSTOR, 10.1198/000313002753631330

Stuart Elizabeth A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” Statistical Science : A Review Journal of the Institute of Mathematical Statistics 25 (1): 1–21. 10.1214/09-STS313. [PubMed: 20871802]

Stuart Elizabeth A, DuGoff Eva, Abrams Michael, Salkever David, and Steinwachs Donald. 2013. “Estimating Causal Effects in Observational Studies Using Electronic Health Data: Challenges and (Some) Solutions.” EGEMS (Washington, DC) 1 (3): 1038. 10.13063/2327-9214.1038.

Susanti Yuliana, Pratiwi Hasih, Sulistijowati Sri H, and Liana Twenty. 2014. “P A M Estimation, S Estimation, And Mm Estimation in Robust Regression.” International Journal of Pure and Applied Mathematics 91 (3): 349–60. 10.12732/ijpam.v91i3.7.

Tian Lu, Zhao Lihui, and Wei LJ. 2014. “Predicting the Restricted Mean Event Time with the Subject’s Baseline Covariates in Survival Analysis.” Biostatistics 15 (2): 222–33. 10.1093/biostatistics/kxt050. [PubMed: 24292992]

Tyree Patrick T, Lind Bonnie K, and Lafferty William E. 2006. “Challenges of Using Medical Insurance Claims Data for Utilization Analysis.” American Journal of Medical Quality : The Official Journal of the American College of Medical Quality 21 (4): 269–75. https://doi.Org/10.1177/1062860606288774. [PubMed: 16849784]

Uno Hajime, Claggett Brian, Tian Lu, Inoue Eisuke, Gallo Paul, Miyata Toshio, Schrag Deborah, et al. 2014. “Moving beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis.” Journal of Clinical Oncology 32 (22): 2380–85. 10.1200/JC0.2014.55.2208. [PubMed: 24982461]

VanderWeele Tyler J. 2019. “Principles of Confounder Selection.” European Journal of Epidemiology 34 (3). 10.1007/s10654-019-00494-6.

Venkatesh Arjun K., Mei Hao, Kocher Keith E., Granovsky Michael, Obermeyer Ziad, Spatz Erica S., Rothenberg Craig, Krumholz Harlan M., and Lin Zhenqui. 2017. “Identification of Emergency Department Visits in Medicare Administrative Claims: Approaches and Implications.” Academic Emergency Medicine 24 (4): 422–31. 10.1111/acem.13140. [PubMed: 27864915]

von Elm Erik; Altman Douglas G.; Egger Matthias; Pocock Stuart J.; Gøtzsche Peter C.; Vandenbroucke Jan P.‖ The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies, Epidemiology: 11 2007 - Volume 18 - Issue 6 - p 800–804 doi:10.1097/EDE.0b013e31815776 [PubMed: 18049194] ‖

Ward E, Jemal A, Cokkinides V, Singh GK, Cardinez C, Ghafoor A and Thun M (2004), Cancer Disparities by Race/Ethnicity and Socioeconomic Status. CA: A Cancer Journal for Clinicians, 54: 78–93. doi:10.3322/canjclin.54.2.78

Van Der Weele Tyler J., and Ding Peng. 2017. “Sensitivity Analysis in Observational Research: Introducing the E-Value.” Annals of Internal Medicine 167 (4): 268–74. 10.7326/M16-2607. [PubMed: 28693043]

Weitzen Sherry, Lapane Kate L., Toledano Alicia Y., Hume Anne L., and Mor Vincent. 2004. “Principles for Modeling Propensity Scores in Medical Research: A Systematic Literature Review.” Pharmacoepidemiology and Drug Safety 13 (12): 841–53. 10.1002/pds.969. [PubMed: 15386709]

Westreich Daniel, Lessler Justin, and Funk Michele Jonsson. 2010. “Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (CART), and Meta-Classifiers as Alternatives to Logistic Regression.” Journal of Clinical Epidemiology 63 (8): 826–33. 10.1016/J.JCLINEPI.2009.11.020. [PubMed: 20630332]

Wilson John, and Bock Adam. 2012. “Optum The Benefit of Using Both Claims Data and Electronic Medical Record Data in Health Care Analysis White Paper.”

Wood Simon, Pya Natalya, and Safken Benjamin. 2018. “Package ‘mgcv’ Title Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.”

Wyss Richard, Ellis Alan R., Brookhart M. Alan, Girman Cynthia J., Funk Michele Jonsson, LoCasale Robert, and Stürmer Til. 2014. “The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, BCART, and the Covariate-Balancing Propensity Score.” American Journal of Epidemiology 180 (6): 645–55. 10.1093/aje/kwu181. [PubMed: 25143475]

Wyss Richard, Girman Cynthia J, LoCasale Robert J, Brookhart Alan M, and Stürmer Til. 2013. “Variable Selection for Propensity Score Models When Estimating Treatment Effects on Multiple Outcomes: A Simulation Study.” Pharmacoepidemiology and Drug Safety 22 (1): 77–85. https://doi.Org/10.1002/pds.3356. [PubMed: 23070806]

Yao Xiaoxin I., Wang Xiaofei, Speicher Paul J., Hwang E. Shelley, Cheng Perry, Harpole David H., Berry Mark F., Schrag Deborah, and Pang Herbert H.. 2017. “Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies.” JNCI: Journal of the National Cancer Institute 109 (8). 10.1093/jnci/djw323.

Yu Bin, and Kumbier Karl. 2020. “Veridical Data Science.” Proceedings of the National Academy of Sciences 117 (8): 201901326. 10.1073/pnas.1901326117.

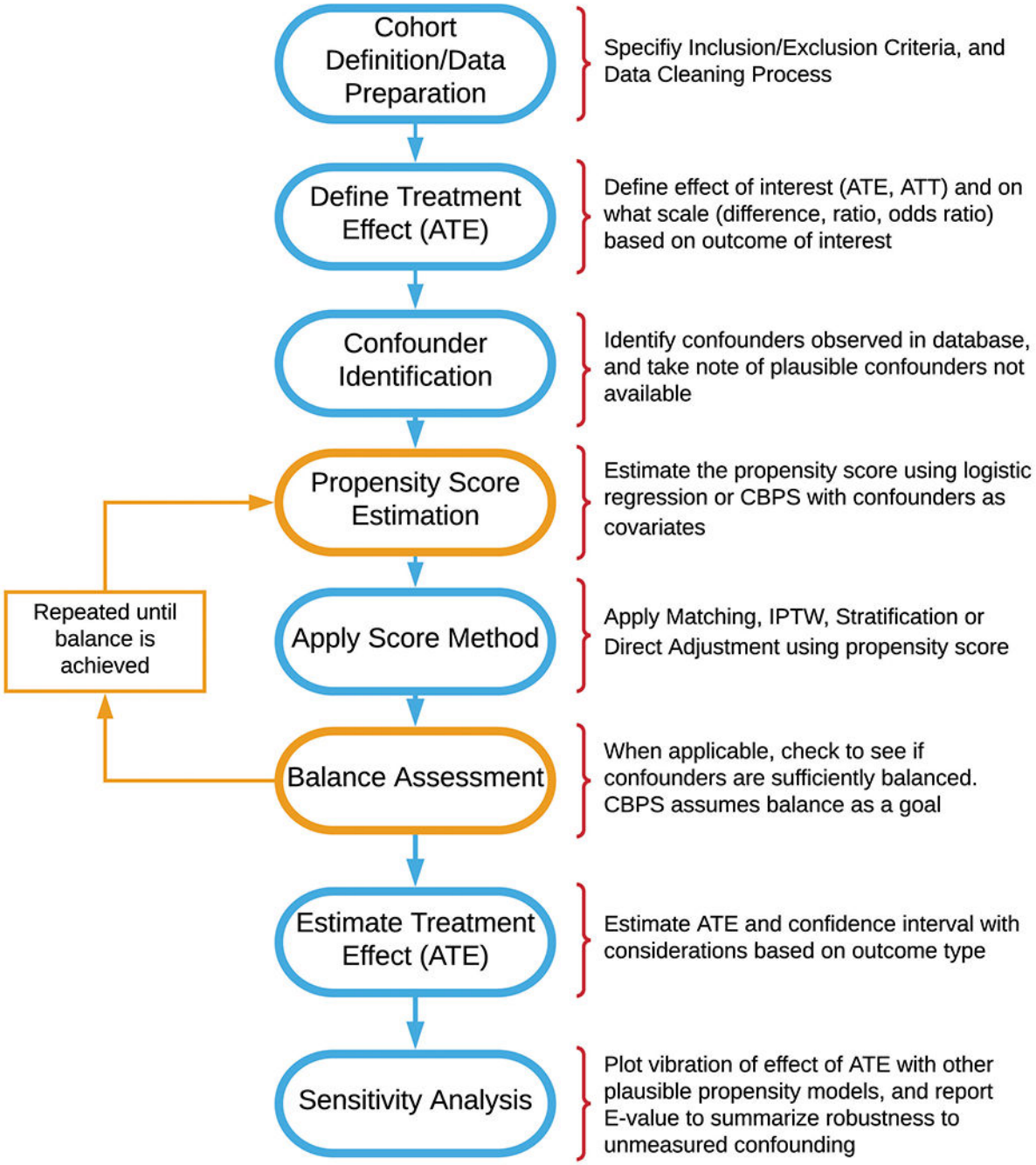

Figure 1:

Comparative Effectiveness Data Analysis Pipeline Flow Diagram.

Gold pathway indicates steps done in iteration until acceptable balance is achieved.

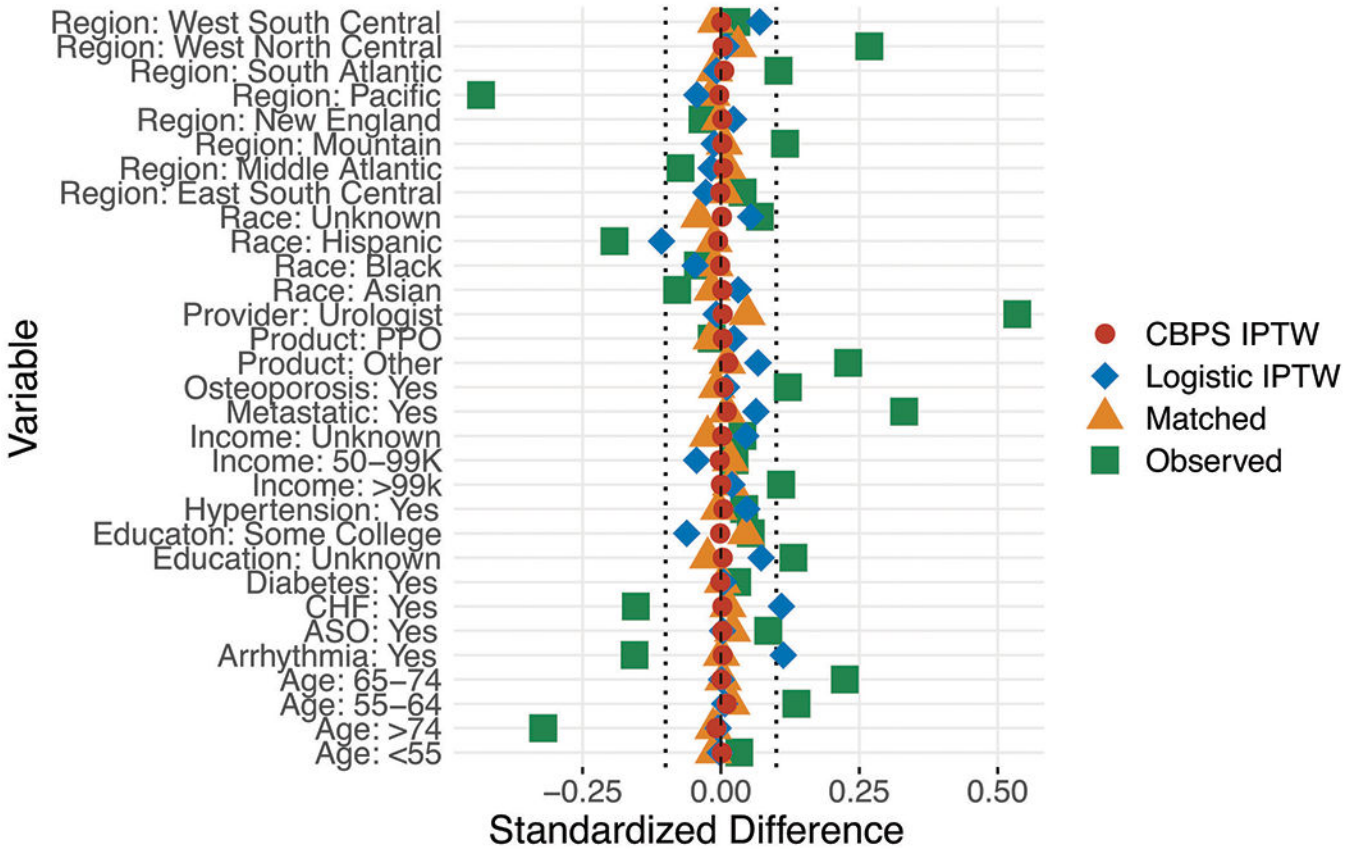

Figure 2:

Balance Diagnostics Plot

Standardized differences shown for each confounder variable. Vertical dotted lines indicate the desired balance level. Differences shown for the observed data, after matching, and weighting with both calculated propensity scores (logistic regression and CBPS)

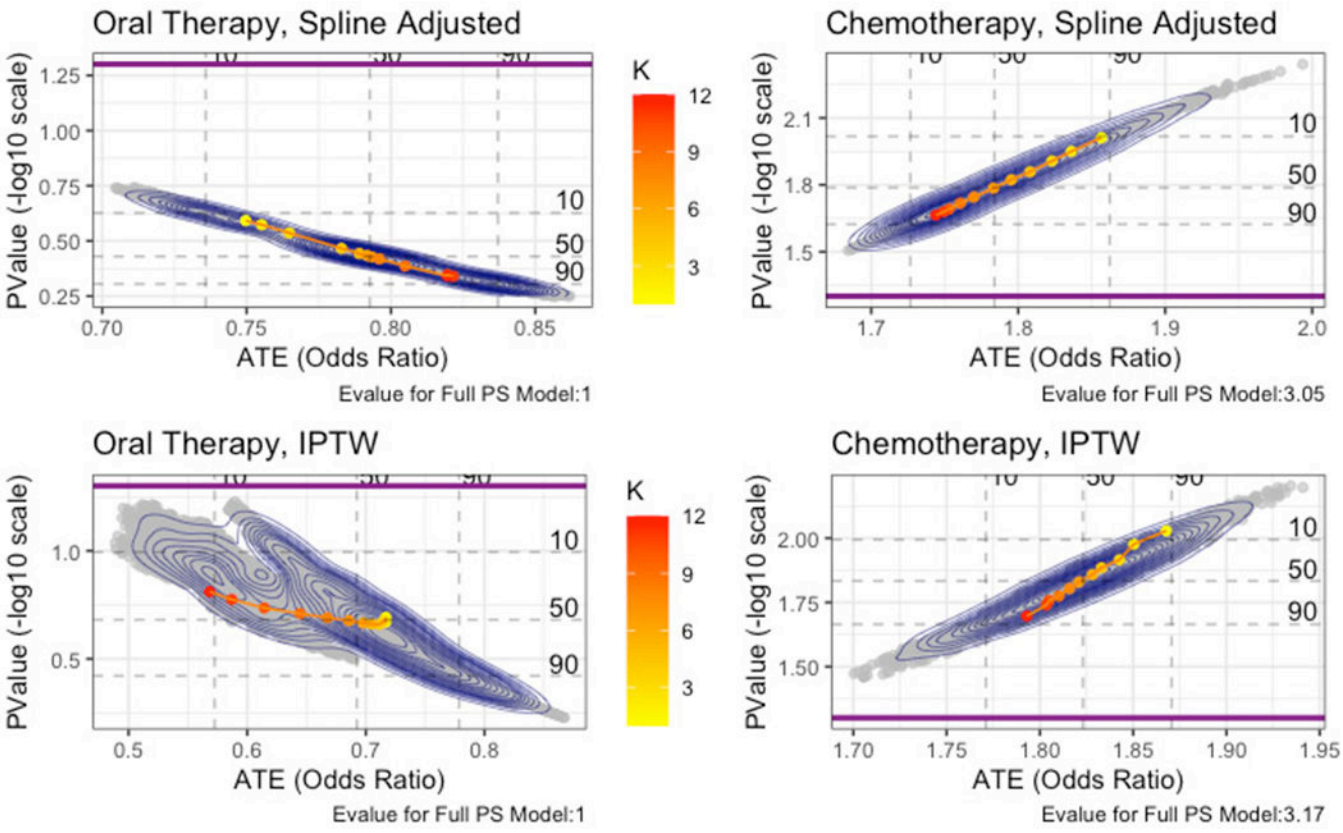

Figure 3:

Visualized Sensitivity Analysis

Four sensitivity analyses for four ATE estimates are shown. Contours over scatterplot show the entire distribution of ATE and associated p-values for the set of plausible propensity score models. Dashed lines show denoted percentiles cutoffs for this distribution. K denotes number of covariates in shown model, and the dotted line plot shows median ATE and p-value for each set of K covariates from K=1 to K=12. Thick solid line indicates significance threshold of alpha=0.05. E-values for the full model (K=12) are listed in caption.

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)
Longitudinally Varying 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[2]	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 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]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)

	Immunotherapy (N = 504)		Chemotherapy (N = 2,214)		Oral Therapy (N = 2,747)	
Variable	**Count**	**(%)**	**Count**	**(%)**	**Count**	**(%)**
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	(5J)	204	(74)

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)**
Longitudinally Varying Outcome: Mean Daily 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 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.