ELSEVIER

Seminars in
RADIATION
ONCOLOGY

# Analytical Methods for Observational Data to Generate Hypotheses and Inform Clinical Decisions

Todd A. DeWees, PhD,* Carlos E. Vargas, MD,† Michael A. Golafshar, MS,*
William Scott Harmsen, MS,‡ and Amylou C. Dueck, PhD*

**Randomized controlled trials have been considered the gold standard in informing clinical decision-making while observational studies have generally been utilized to generate hypotheses for future studies. The rising cost of randomized studies along with increased difficulty in accrual has led the clinical community to consider utilizing observational studies to inform clinical decisions. Various statistical methods exist to analyze observational data. Researchers must consider each method carefully, paying specific attention to its ability to answer the hypotheses, while ensuring the underlying assumptions are met. While each has its own strengths and weaknesses, research has shown that each method may yield similar estimates of treatment effect when conducted appropriately. We describe several commonly used analytical methods including their: strengths, weaknesses, and common missteps in order to inform and serve as a reference to the broader oncology community.**

**Semin Radiat Oncol 29:311−317 © 2019 Elsevier Inc. All rights reserved.**

## Introduction

In oncology overall and radiation oncology specifically, there are several well-recognized challenges in conducting randomized control trials (RCTs) to answer clinically relevant questions, including: long study time and high costs, patient unwillingness to be randomized, and lack of generalizability in RCT results. Furthermore, some research questions cannot be ethically answered utilizing RCT.[1] The challenges associated with developing and carrying out RCTs have led to the use observational studies.

Studies utilizing retrospectively collected institutional data, an example of observational studies, are often construed solely as hypothesis generating. However, meta-analyses comparing observational studies and RCT for the same populations have shown that estimates obtained from observational studies are often within the 95% confidence intervals obtained from the RCT.[2-4] Observational studies and RCT may produce conflicting results due to many factors including selection bias, issues with generalizability, inadequate statistical power, or differing treatment and/or follow-up approaches.[5]

Differing types of observational studies are utilized in research, with varying ability to answer questions and potential bias inherent in each study. Table 1 summarizes the advantages and disadvantages associated with each observational studies design. Planning, interpreting, and proper reporting of analyses for each design are essential, with careful consideration for uncontrolled bias and missing data.

Multiple ways exist to analyze observational data each providing variations in hypothesis construction, methodology implementation, and clinical interpretation of the results.[6] We describe below stratification, multivariate modeling, K:1 matching, propensity score (PS) matching, PS adjustment, and instrumental variables (IVs). Alternate methods scarcely used in the medical literature will be briefly introduced where appropriate. Table 2 describes strengths and weakness of methods most commonly used in the radiation oncology literature.

## Statistical Analysis of Observational Studies

### Stratification

Stratification is often the initial approach utilized by researchers in observational studies. Subgroups are created

*Division of Biomedical Statistics and Informatics, Mayo Clinic, Scottsdale, AZ
†Department of Radiation Oncology, Mayo Clinic, Phoenix, AZ
‡Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Conflicts of Interest: None.

Address reprint requests to Todd A. DeWees, PhD, Division of Health Sciences Research, Mayo Clinic, 13400 E. Shea Blvd., Scottsdale, AZ 85259. E-mail: DeWees.Todd@mayo.edu

https://doi.org/10.1016/j.semradonc.2019.05.003
1053-4296/© 2019 Elsevier Inc. All rights reserved.

**Table 1** Strengths, Weaknesses, and Best Clinical Application for RCT and Observational Studies

| Types of Study | Description | Advantages | Disadvantages | Best Application |
|---|---|---|---|---|
| RCT | Interventional: with subjects randomized to treatment vs control | Structured inclusion criteria minimizing bias, standard analytic methods, identifies "causality" | Cost and time. Limited generalizability, often unable to detect small effects or conduct subset analyses. | Phase III with specific hypotheses and the desire to alter clinical practice |
| Case series | Observational: Descriptive with the intent to identify new, unobserved findings and generate hypotheses | Rapidly bring attention to new findings for future research, low cost and time | Low generalizability, definite conclusions cannot be made, potential for bias/confounding | Preliminary data for Phase I/II trials and to generate hypotheses for preliminary studies |
| Cross sectional | Observational: assesses exposure and disease prevalence at a specific time point | Ability to analyze relationship between multiple exposures and diseases, straightforward subject selection that can be used to be generalized to populations | Cannot measure incidence rates or risk and unable to determine causality or temporality. Potential for bias/confounding | Best used for quick assessment of prevalence to generate hypotheses for future studies. |
| Case control | Observational: utilizes real-world clinically diverse population to compare disease cases with control patients for outcomes based on exposure status. | Generally shorter observation periods and lower costs compared to cohort or RCT. No patient follow-up needed, generalizable based on selection of cases and controls. | Potential for bias/confounding as choice of cases and controls need to be valid and generalizable | Analysis involving cases and controls with common exposure and a single disease status of interest. Retrospectively treated patients with rare diseases. |
| Prospective cohort | Observational: patients are followed over time to assess disease development following exposure. | Ability to utilize real-world clinical populations with high generalizability. Can directly measure incidence and risk on multiple outcomes. Generally less measurement error than alternative observational studies | Generally requires larger sample size and longer follow-up that can be more expensive than other observational studies. Difficult to find causality. Careful consideration of measured and unmeasured bias/confounding needed | Best to assess the effect of exposure on possibly multiple outcomes of interest, with the desire of estimating risk based on exposure. |

**Table 2** Strengths and Weaknesses of Different Analytic Techniques for Observational Studies

| Method | Strengths | Weaknesses |
|---|---|---|
| Stratification | Simple with clear interpretation<br>No assumptions necessary<br>Ability to see effect modification | Only few categories or covariates can be used<br>Requires understanding of confounder covariate structure |
| Multivariate analysis (MVA) | Can include many confounders<br>Can examine effect of confounders<br>Ability to examine multilevel effects<br>Utilizes all subjects | Focus is not set on balancing data between treatment groups<br>Adequate overlap between groups is difficult to assess<br>Potentially poor model fit |
| K:1 matching | Simple<br>Balances confounding factors | Difficult to find matches, possible reduction in sample size<br>Need to weigh effect of under- and overmatching<br>Unable to examine the effect of the confounders used in matching on outcomes |
| Propensity score (PS) matching | Single number generated for simple matching<br>Ability to assess bias between groups<br>Many covariates are possible | Potentially matching very different patients with similar scores<br>Performs better with only a few number of events per confounder<br>Possible reduction in sample size |
| Propensity score (PS) adjustment | Many covariates are possible<br>Bias due to misspecification of the model is less than MVA | Improper model selection may lead to biased weights<br>May need to remove some patients with outlying weights |
| Instrumental variables (IV) | Only a single variable is needed<br>Utilizes all subjects<br>Ability to address questions where other types of adjustment cannot be completed.<br>The only method that can mitigate unobserved confounding | Difficult to identify IV<br>Difficulty ensuring IV is not at all directly associated with outcome |

based on clinical knowledge (eg, pediatrics vs adults) and treatment effect is evaluated separately on those subgroups. The main advantage is obtaining intuitive, interpretable results giving insight into the balanced/imbalanced nature of the data and the effect on the treatment groups, while not requiring assumptions of the underlying relationship between the stratification variable and the outcome. The need to limit strata to ensure analyzable subgroups is an important consideration.[7] For instance, in assessing the effect of hypofractionation vs standard fractionation on lung cancer a researcher may want to consider three age strata (<60, 60-80, >80), 3 smoking status strata (never, past, and current), and 2-stage strata (<III vs III+), resulting in 18 subgroups, some with few patients and nontrivial analysis and interpretation. Alternative techniques such as asymmetric stratification exist in the attempt to increase the number of covariates that can be analyzed within stratification, but are rarely utilized in medical literature.

Stratification can be used to inform alternate methods of analysis: matching, multivariable (MVA) analysis, or PS analysis. Alternatively, stratification of retrospective data is useful to inform prospective stratification in future RCT.

## Multivariable Analysis

A univariate model can be used to investigate the impact of an individual covariate on an outcome, for instance a Cox regression model of local recurrence based on type of treatment received. Models directly testing treatment effect without regard to possible confounders are termed "unadjusted." To investigate the impact of multiple covariates, 2 general methods of MVA modeling are utilized. In "adjusted" MVA, the researcher utilizes historical/known confounders as a priori covariates in an attempt to reduce bias between treatment groups. The treatment variable is then added to determine independent treatment effect once the known confounders are controlled for in the model. Interpretation and *P* values are inappropriate for variables utilized as controls for confounding/bias.

MVA can also be used to build a model from covariates that have already been shown to be clinically and statistically significant in univariate model. Multiple statistical methods for selecting among covariates exist (eg, backwards, forwards, and stepwise). Each of these methods has advantages and disadvantages that must be considered throughout the model building process. The effect of inclusion or omission of a covariate on the parameter estimates for all other covariates must be carefully inspected. Another important issue is multicollinearity; that is, when explanatory covariates are correlated with each other.[8]

The advantage of MVA is its ability to incorporate a variety of covariate (categorical and continuous) and endpoint distributions (binomial, nominal, ordinal, quantitative, and censored data). General linear mixed models (eg, quality of life, tumor growth), logistic regression (eg, tumor response, necrosis), and Cox proportional hazards (PH) or competing risks models (eg, overall survival, local control) are generally used in the medical literature. Straightforward and interpretable results can be obtained to inform clinical decision-making through slopes (linear regression), odds ratios (logistic regression), and hazard ratios (Cox PH).

Another option in MVA is a severity of illness score (or other single covariate utilized to summarize multiple clinical indicators). This covariate is ideally developed and validated on different groups of patients including those of interest to the proposed study (eg, Charlson Comorbidity Index). While using such covariates generally leads to proper adjustment of confounding (ie, similar treatment effect point estimates), the statistical significance of the covariate itself is often exaggerated.[9]

In MVA, a model captures the relationship between covariates and the outcome of interest. There is no explicit balancing of groups, as in matching methods, thus the difficulty in directly comparing and assessing overlap between groups can be a disadvantage.

A significant assumption in medical publications is constant effect over time (ie, PH). Inclusion of covariates that may affect the outcome differently over time (eg, late toxicity affecting long-term survivors) must be taken into account and modeled appropriately. Model fit (eg, $R^2$) and model discrimination (eg, Hosmer-Lemshow c-statistic, receiver operating characteristic Curves) are essential to ensure that models are not over- or underfitting the data.[10] When implementing predictive modeling, external and internal validation techniques (eg, bootstrapping) should be implemented to evaluate the model performance in diagnostic and prognostic settings.[11]

## K:1 Covariate Matching

Matching can be employed prospectively on subjects prior to the observance of outcome or retrospectively on subjects after all data are known. A researcher must ensure that the outcome values are not utilized in the matching process resulting in a sample that leads to a desired result.[6] There are 4 main steps to implementing matched analysis: (1) define the closeness of your match (exact, interval, and calibration metric); (2) implement the matching method based on closeness; (3) iteratively select matched samples assessing the quality of the match at each iteration; and (4) analyze the treatment effect for the outcome based on the matched samples.[12] Closeness is important as a researcher may desire exact matching (all covariates are equal), interval matching (eg, age within 5 years), or the creation of a single distance measure of combined covariate differences. Closeness and quantity of matched covariates affect the number and balance of the remaining matched subjects. Too much distance or too few covariates may result in poor matching or an increase in unexplained confounding; conversely, the inclusion of too many unassociated covariates may also result in an increase in unexplained confounding. Researchers should utilize covariates known to be related to the outcome of interest, complete the matching, and check the balance. It is advisable to choose matching variables based on previous research and scientific understanding, without the knowledge of observed outcomes. Similarly, choosing matching variables that may have been affected by the treatment[13] or variables that are predictive of treatment assignment hinder the ability to detect true treatment effect.

Once covariates and distance measures are selected, the ratio of control:case matched pairs (1:1 or K:1) and the sample sizes needed to achieve balanced matching must be determined. A more precise match, larger number of matching covariates, and higher matched ratios exponentially increase the number of control subjects needed. The matching ratio often creates a variance trade-off. Increasing the ratio of controls decreases overall variance via a larger matched cohort offering more statistical information, but often increases the between-matched pair variance by adding more matches that are the second or higher closest match.[14]

The quality and balance of the matched pairs should be assessed. Overlap, in terms of matching variables, between controls and cases may be minimal, making the matched analysis inappropriate. Conversely, relaxing matching criteria in the setting of highly heterogenous groups will not provide reliable treatment comparisons. For example, when matching surgery to RT patients, the characteristics may be so different between groups that matching is difficult to impossible. To assess balance in all study covariates between matched pairs, 2 group tests (eg, Wilcoxon and Fisher's Exact tests) are generally acceptable assuming samples are drawn without replacement.[15]

## Propensity Scores

PSs[16] are an alternative way to adjust for confounding in nonrandomized observational studies, by calculating the conditional probability of a subject receiving treatment in order to reduce bias in the analysis of a treatment effect. PS are often thought of as a method for creating pseudo-RCT by utilizing logistic regression to predict treatment assignment with pretreatment/baseline covariates deemed to be confounders. For example to compare disease outcomes (eg, overall survival) for proton therapy vs photon therapy, an initial logistic model would use patient baseline characteristics to predict treatment assignment. The treatment assignment probabilities from that logistic model would then be used to either match patients, create inverse weights, or create a covariate to be used in a subsequent regression model for PS adjustment in order to compare the effect of proton therapy vs photon therapy for disease outcome. The resulting PS is then incorporated into statistical analyses as described in subsequent sections.

The success of PS modeling is evaluated by comparing the balance between treatment and control groups (eg, those receiving proton therapy vs those receiving photon therapy) after accounting for the PS through matching or score adjustment. Variable selection is of utmost importance in computing PS, having been selected a priori based on clinical knowledge and previous literature. Having many predictors is often better than too few, and since the goal is prediction, multicollinearity among predictors is not of concern.[6,12] It is necessary to ensure that variables weakly associated with treatment but strongly associated with outcome are not removed. Commonly used methods for assessing and reporting PS model fit are c-statistics and absolute standardized differences. As with any other model fit analysis, overall balance should be inspected before progressing to treatment effect analysis.[13]

PS models require all confounders be measured. If unmeasured confounders exist (eg, The Surveillance, Epidemiology, and End Results (SEER) Program does not contain smoking information) then residual confounding may exist that cannot be addressed. Studies demonstrating differing results between RCT and PS adjusted observational studies are often due to a lack of control for unmeasured confounders. Heterogenous PS distributions between treatment groups may necessitate multiple strategies. "Trimming" (ie, removal) of patients to create overlapping PS, reduces the sample size and creates a loss of power and potential bias.[17] Nonoverlapping PS would indicate patient groups were highly dissimilar and would have been highly unlikely to have been candidates for the other treatment arm. For example, in comparing surgery to radiotherapy, patients with higher levels of comorbidities may have never been candidates for surgery and therefore have very small PSs that do not allow for matching or weighting, necessitating their being "trimmed" from the analysis.

## PSs Matching

Once PS is calculated, it can be used to match cases and controls with similar scores. Matching methods include variations on: distance measures, determining the threshold for score similarity, the ratio of matched patients, and whether or not to match the same control subject to multiple treatment subjects (ie, drawing with replacement).[18] Distance is optimally calculated by utilizing the difference between scores on the logit scale. The threshold for score similarity is often referred to as the "caliper." Narrow calipers may result in many subjects going unmatched. Large calipers include more subjects, but may result in imbalance. Appropriate caliper size depends on multiple factors relative to the variations among the treatment group PS. The ratio of controls to cases as well as sampling with replacement should be considered similarly to previously described section on covariate matching, and considerations should be made in assessing the trade-offs between increasing sample size and reducing variability. Advantages and disadvantages of PS matching are similar to considerations in the previous section on covariate matching.

## PS Adjustment

Two types of PS models are typically found in literature. PS covariate adjustment involves use of MVA analysis, by adjusting for baseline confounding through the PS as a covariate. This method uses all data, but has several disadvantages including: likely model misspecification; produces only treatment effect odds ratios or hazard ratios and not absolute treatment effect, average treatment effect, or average treatment effect for those treated due to score adjustments altering underlying averages; shown to increase bias in ratio estimates; and difficulty in assessing balance. The PS adjustment method is not ideal and should typically not be utilized.

Inverse probability of treatment weighting (IPTW), in which a synthetic sample is created such that the distribution of baseline covariates involved in the PS is independent of treatment assignment,[19] is an often used method. The inverse of a function of the PS is used as a weight in weighted regression analyses. Thus, subjects who are less likely to receive treatment but actually got treatment would receive a higher weight, while control patients who are less likely to receive treatment receive a lower weight.[19] Like PS covariate adjustment methods, inverse probability of treatment weighting allows for use of all data but also allows for estimation of both average treatment effect, average treatment effect for those treated, and reduces bias.[20] A disadvantage to this method is that the misspecification of the PS model may have significant impact on the weights and therefore the treatment effect estimates.

## Instrumental Variables

The main advantage to IV analyses is that unlike other methods that can only control for observed confounders, instruments can also potentially control for unobserved confounders. However, this method is uncommonly used in the medical literature because finding an appropriate IV is often difficult.

Three key assumptions of IV need to be met: IV should be associated with treatment due to a common cause; IVs are independent of other patient characteristics; and IV cannot be directly associated with outcome, except through its association with treatment.[21] Instruments common in literature are patient's distance to care, facility/physician variation (practice variation by site/physician), and treatment cost to patient after insurance (eg, cost to patient for proton therapy).

Pollon et al utilized an IV in their analysis on the effect of intensity-modulated radiation therapy (IMRT) on hospitalization for patients with anal squamous cell carcinoma.[22] They chose to utilize "physician affinity" for ordering IMRT for treating nonanal gastrointestinal cancers as their IV as it would be associated with IMRT, but less likely to be associated with factors correlated to the treatment of anal cancer or directly associated with treatment outcomes.

**Table 3** Steps for Planning, Conducting, and Interpreting Observational Studies

| | |
|---|---|
| Planning | |
| Step 1 | Ensure the database and study population fit the research question |
| Step 2 | Ensure the size of the study population is large enough to answer the question |
| Step 3 | Ensure the study design provides answers for clinical decision-making (eg, cohort studies for assessing risk and rates) |
| Step 4 | Ensure that the treatment, outcome, and potential confounding covariates are accurately measured with long enough follow-up to answer the research question |
| Step 5 | Ensure the statistical methodology is capable of answering the research question while meeting all necessary assumptions |
| Reporting | |
| Step 1 | The original sample sizes and their characteristics for all patients as well as the number of analyzed patients in the final model are reported |
| Step 2 | A description of the variables used for inclusion in either the MVA, PS, IV, or other matching method utilized along with the specified regression model used |
| Step 3 | The type of matching algorithm used along with a description of the "distance" metric |
| Step 4 | Diagnostics demonstrating the quality of the resulting match (cohort balance) |
| Step 5 | An indication of the type of statistical methods utilized |
| Step 6 | Sensitivity analysis to show the robustness of the determined treatment effect to variations in unmeasured confounding and statistical models |

# Pitfalls and Reporting

## Missing Data

A significant area of study is dealing with missing data in explanatory covariates and outcome. Types of missingness from least to most problematic are: missing completely at random, missing at random, and missing not at random (MNAR). The first 2 can be remedied utilizing a variety of approaches such as multiple imputation where multiple data sets are created (each with missing data filled in with a new set of likely data values) and analyzed with results taken from pooled estimates.[23] However, if missingness falls under MNAR then alternative methods may be necessary. One way in which MNAR occurs in oncology is if the effects of treatment led to an acute hospitalization which prevents data collection at a planned study time point (eg, completion of patient questionnaires, blood draws, scans, etc).

## Sensitivity Analysis

Sensitivity analyses aid in ensuring that results remain consistent across a variety of assumptions. This is particularly salient in observational data analysis as many assumptions are untestable. Sensitivity analyses can be used to investigate the likelihood of unidentified or unknown confounders in modeling, and attempt to create bounds for assessing the impact of an unmeasured confounder after adjusting for measured covariates. For example, Lin et al[24] demonstrated that the plausibility of treatment effect from observational studies increases if the effect is consistent across plausible assumptions of unmeasured covariates. Therefore, one approach to sensitivity analyses is to utilize each of the previously described methods (eg, matching, MVA, and PS), compare results, and determine if different methodologies result in differing estimates of the treatment effect.[25]

## Interpretation

All study designs have flaws that can threaten external (generalizability of results) and internal (improper/missing data) validity. However, if guidelines are followed when planning, analyzing, and interpreting observational studies for published literature then readers will be able to incorporate the data into their clinical decision-making.[18,26] Table 3 provides the necessary steps for conducting and reporting observational research. Utilizing these techniques for planning and reporting of observational studies, the researcher can present their own interpretation of the results, while allowing the reader to effectively create their own understanding of the research question. STROBE[27] has created extensive guidelines on the conduct and dissemination of observational studies with the aim of STrengthening the Reporting of OBservational studies in Epidemiology.

# Conclusions

While RCTs are the gold standard among research designs, alternative observational designs and approaches permit strong causal inference in settings where RCTs are not feasible. Prospective observational studies can enroll an entire patient population without or with limited exclusion criteria (eg, pragmatic trials) enabling larger population sizes that are generally less costly and faster to conduct than RCT.

Several studies have shown observational studies have demonstrated similar results to that of RCT.[4,28] Similarly, multiple publications have demonstrated the similarities in estimated treatment effects between MVA and PS,[29,30] PS and IV,[31,32] and comparisons of all 3 methods[33]; including within radiation oncology for patients diagnosed with anal squamous cell carcinoma based on the SEER database.[22,34,35]

Understanding the underlying biases from unmeasured confounders is paramount to the utilization and interpretation to observational studies. Great care is needed to ensure that the methods being utilized are transparent.[13,36] While observational studies are criticized for the likelihood of bias and confounding, steps can be taken to ensure proper study design, properly collected data, and analyses that are able to test the research hypotheses and thereby inform clinical decisions.

## References

1. Nardini C: The ethics of clinical trials. Ecancermedicalscience 8:387, 2014
2. Ioannidis JP, Haidich AB, Pappa M, et al: Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 286:821-830, 2001
3. Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342:1887-1892, 2000
4. Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. N Engl J Med 342:1878-1886, 2000
5. Black N: Why we need observational studies to evaluate the effectiveness of health care. Br Med J 312:1215-1218, 1996
6. Rubin DB: The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Stat Med 26:20-36, 2007
7. Rubin DB: Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757-763, 1997
8. Harrell FE: Regression Modeling Strategies. With applications to Linear Models, Logistic Regression, and Survival Analysis. Springer Verlag; 2002
9. Miettinen OS: Stratification by a multivariate confounder score. Am J Epidemiol 104:609-620, 1976
10. Angus DC, Pinsky MR: Risk prediction: Judging the judges. Intensive Care Med 23:363-365, 1997
11. Han K, Song K, Choi BW: How to develop, validate, and compare clinical prediction models involving radiological parameters: Study design and statistical methods. Korean J Radiol 17:339-350, 2016
12. Stuart EA: Matching methods for causal inference: A review and a look forward. Stat Sci 25:1-21, 2010
13. Brookhart MA, Schneeweiss S, Rothman KJ, et al: Variable selection for propensity score models. Am J Epidemiol 163:1149-1156, 2006
14. Rubin DB, Thomas N: Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc 95:573-585, 2000
15. Hansen BB: The essential role of balance tests in propensity-matched observational studies: Comments on ‘A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003’ by Peter Austin, Statistics in Medicine. Stat Med 27:2050-2054, 2008
16. D’Agostino AR Jr: Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 17:2265-2281, 1998
17. Streiner DL, Norman GR: The pros and cons of propensity scores. Chest 142:1380-1382, 2012
18. McMurry TL, Hu YN, Blackstone EH, et al: Propensity scores: Methods, considerations, and applications in the Journal of Thoracic and Cardiovascular Surgery. J Thorac Cardiov Sur 150:14-19, 2015
19. Austin PC: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 46:399-424, 2011
20. Austin PC: The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 26:3078-3094, 2007
21. Brookhart MA, Rassen JA, Schneeweiss S: Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidem Dr S 19:537-554, 2010
22. Pollom EL, Wang GY, Harris JP, et al: The impact of intensity modulated radiation therapy on hospitalization outcomes in the SEER-medicare population with anal squamous cell carcinoma. Int J Radiat Oncol 98:177-185, 2017
23. Schafer JL: Analysis of Incomplete Multivariate Data. London, United Kingdom: Chapman & Hall/CRC Press; 1997
24. Lin DY, Psaty BM, Kronmal RA: Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54:948-963, 1998
25. D’Agostino RB, D’Agostino RB: Estimating treatment effects using observational data. JAMA 297:314-316, 2007
26. Sorensen HT, Lash TL, Rothman KJ: Beyond randomized controlled trials: A critical comparison of trials with nonrandomized studies. Hepatology 44:1075-1082, 2006
27. von Elm E, Altman DG, Egger M, et al: The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Int J Surg 12:1495-1499, 2014
28. Anglemyer A, Horvath HT, Bero L: Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 4:1-46, 2014
29. Biondi-Zoccai G, Marullo AGM, Peruzzi M, et al: Last nail in the coffin for propensity scores in observational cardiovascular studies? J Am Coll Cardiol 69:2575-2576, 2017
30. Shah BR, Laupacis A, Hux JE, Austin PC: Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 58:550-559, 2005
31. Fang G, Brooks JM, Chrischilles EA: Apples and oranges? Interpretations of risk adjustment and instrumental variable estimates of intended treatment effects using observational data. Am J Epidemiol 175:60-65, 2012
32. Laborde-Casterot H, Agrinier N, Thilly N: Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: A systematic review. J Clin Epidemiol 68:1232-1240, 2015
33. Stukel TA, Fisher ES, Wennberg DE, et al: Analysis of observational studies in the presence of treatment selection bias — Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 297:278-285, 2007
34. Chin AL, Pollom EL, Qian YS, et al: Impact of intensity-modulated radiotherapy on health care costs of patients with anal squamous cell carcinoma. J Oncol Pract 13:E992-E1001, 2017
35. Kim E, Kim JS, Choi M, et al: Conditional survival in anal carcinoma using the national population-based survey of epidemiology and end results database (1988-2012). Dis Colon Rectum 59:291-298, 2016
36. King G, Nielsen R: Why propensity scores should not be used for matching. Polit Anal 390:1-20, 2019