carelon

White paper

Designing real-world evidence studies for causal inference

A comprehensive guide to generating stronger and more powerful evidence

This white paper seeks to increase understanding of the factors that inform the design and conduct of causal studies in real-world settings. Authors describe in greater detail how the design of randomized controlled trials (RCTs) promotes causal inference and how 'target trial' thinking can support the use of real-world evidence (RWE) studies to estimate causal effects. Also included are specific considerations for the design of rigorous causal RWE studies as well as selected Carelon Research case studies that solve common RWE design challenges.

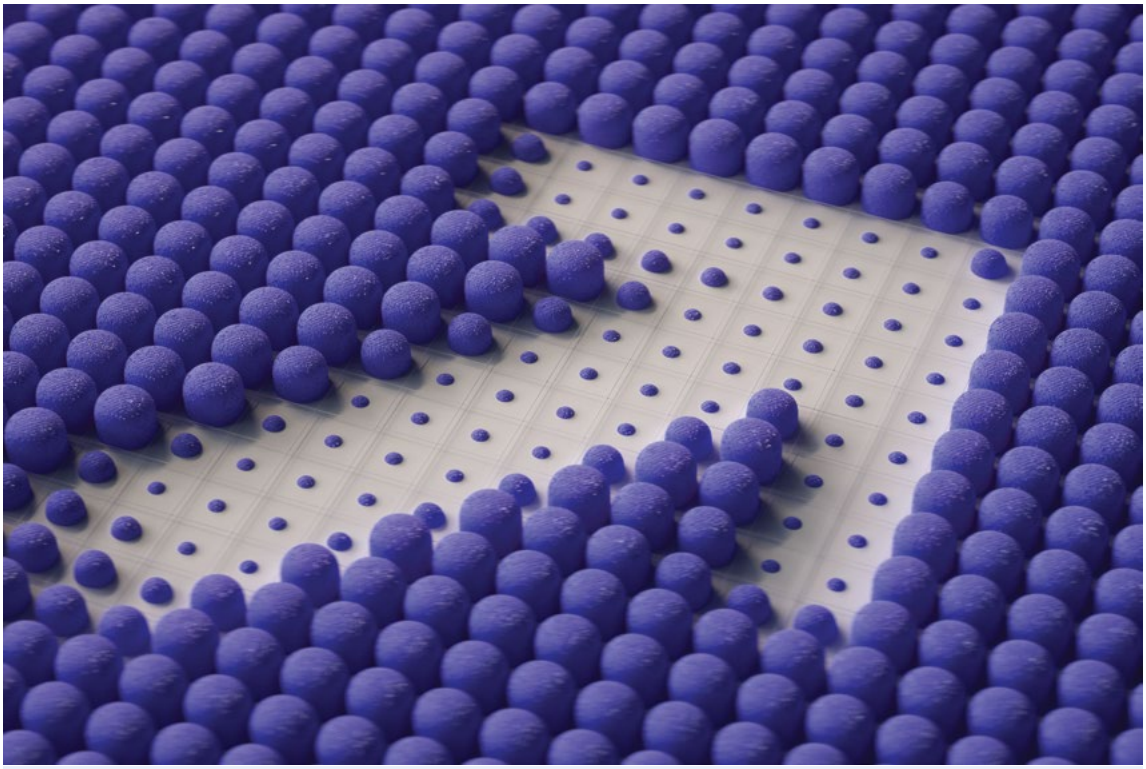


Table of contents

Real-world studies complement randomized controlled trials

The gold standard

RCTs are considered the gold standard for identifying causal relationships between exposure to clinical interventions and hypothesized outcomes due to strong internal validity. RCTs use randomization of study subjects to induce an approximately equal distribution of potentially confounding factors across treatment and control groups. RCTs allow investigators to directly control the timing, dosage, and duration of the intervention and the measurement of endpoints.

For these reasons, high-quality RCTs are commonly considered the preferred approach for measuring the causal effects of medical treatments subject to regulatory approval and evidence-based health insurance coverage decisions.

Limitations of the gold standard

At the same time, RCTs may also suffer from sources of bias, such as cross-over and differential attrition across arms, and limited generalizability. They may be impractical due to high cost, time, and ethical concerns. These limitations of RCTs have stimulated interest over the past several decades in using observational data collected in real-world settings to estimate causal relationships between healthcare interventions and health-related outcomes in a timely, efficient, and externally valid way.

The potential of RWE studies

In many situations, information about causal relationships can be obtained from observational healthcare data, such as administrative data from health insurance plans, national health surveys, and public health surveillance programs. Such data can be used to generate rigorous evidence about causation in the real-world circumstances in which interventions will be used and inform economic and delivery system factors that influence effectiveness in real-world settings. This approach is growing in importance, driven by the need for generalizable and rapidly delivered RWE to inform regulatory, payer, and patient/provider decision making.

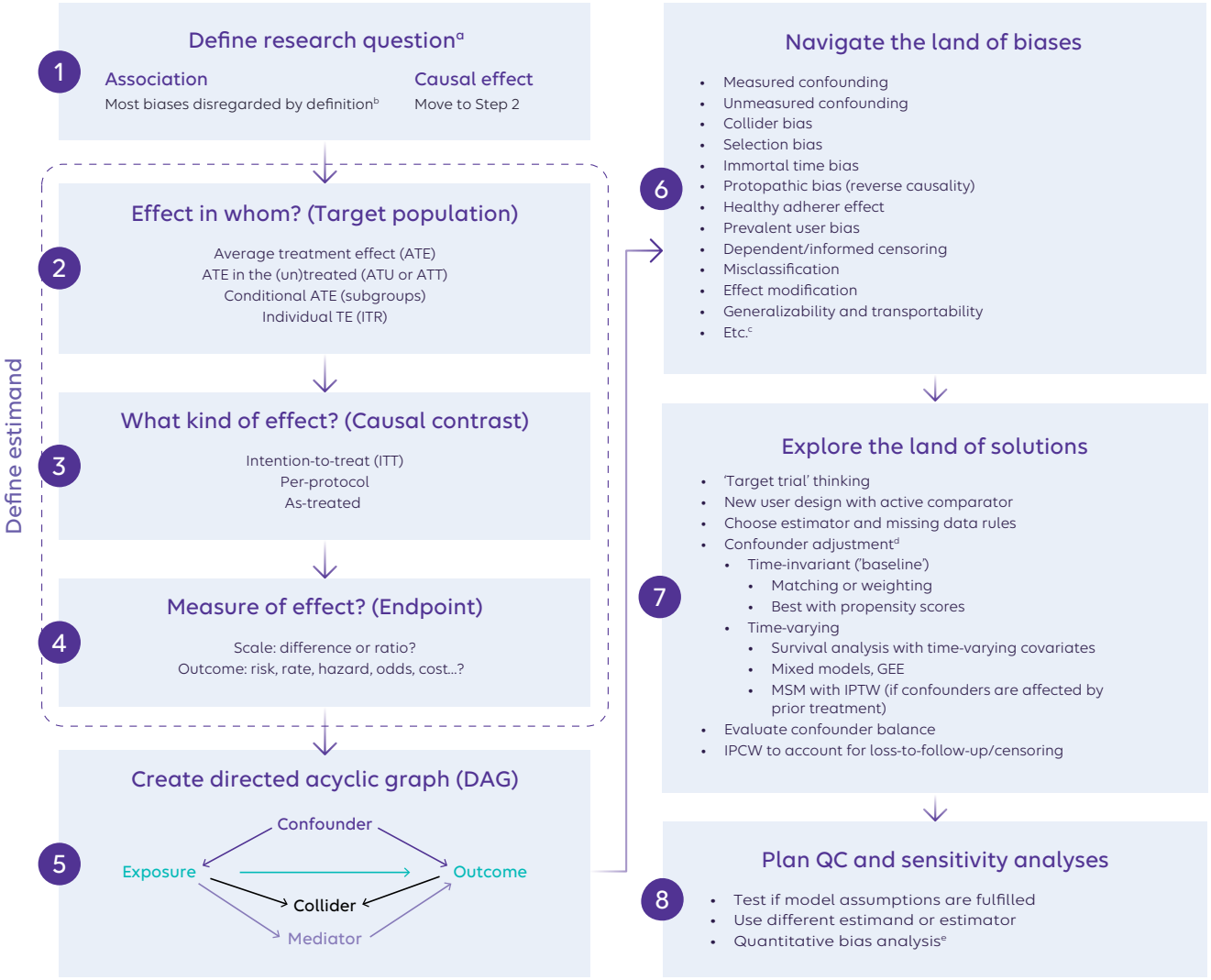
The ‘target trial’ approach

In order to apply RCT-based causal inference best practices to RWE studies, it can be valuable to design a hypothetical RCT protocol or ‘target trial’ that describes eligibility criteria, treatment administration, treatment assignment, specification of outcome measures, length of follow-up, causes of study attrition, causal contrasts (e.g., intent-to-treat or per protocol), and statistical estimation methods.[9-12] Doing so promotes understanding of and alignment between the study outcome of interest and the design and implementation of appropriate real-world studies, which reduces the potential for common biases.

Considerations when designing real-world evidence studies to investigate causal relationships

Causal inference with observational healthcare data combines numerous theoretical and technical concepts; the existing methodological literature on this topic is rich, but can be complex and daunting. Carelon Research developed a step-by-step guide for causal study design that identifies and describes key conceptual issues of importance to researchers designing causal inference studies.

A step-by-step guide to causal study design



Acronyms: GEE, generalized estimating equations; IPC/TW, inverse probability of censoring/treatment weighting; ITR, individual treatment response; MSM, marginal structural model; TE, treatment effect

a. Ensure that the exposure and outcome are well-defined based on literature and expert opinion.
b. More specifically, measures of association are not affected by issues such as confounding and selection bias because they do not intend to isolate and quantify a single causal pathway. However, information bias (e.g., variable misclassification) can negatively affect association estimates, and association estimates remain subject to random variability (and are hence reported with confidence intervals).
c. This list is not exhaustive; it focuses on frequently encountered biases.
d. Only a selection of the most popular pharmacoepidemiological approaches is presented here. Other methods exist, e.g., g-computation and g-estimation for both time-invariant and time-varying analysis, instrumental variables, and doubly robust estimation methods. Program evaluation methods (e.g., difference-in-differences, regression discontinuities) can also be applied to causal questions in healthcare.
e. Online tools include, among others, an E-value calculator for unmeasured confounding (evalue-calculator.com) and the P95 outcome misclassification estimator (apps.p-95.com/ISPE).

Step 1

Define the research question in causal terms

If the research question is explicitly or implicitly causal rather than associative, observational RWE studies can and should be designed to provide such a causal estimate.[4] A causal question should be explicit and easily identified using phrases such as "what is the effect of A on Y?", "how does B affect Y?", and "how does C influence Y?" More implicit causal questions might ask, "how would outcome Y be different, given that exposure B was different from the observed case?" Note that comparative effectiveness/safety studies ask a causal question by definition.

Steps 2 3 4

Estimand vs. estimator

The *estimand* is the causal effect of interest and is described in terms of required design elements: the target population for the counterfactual contrast (e.g., the Average Treatment Effect in the at-risk community)*, the kind of effect (e.g., intent-to- treat or per-protocol, representing different approaches to address changes to treatment during follow-up), and the effect/outcome measure (e.g., survival rates).† This concept is distinct from that of an *estimator*, which is a method of analysis to compute an estimate (numerical value) of the estimand using the available data.[13-16]

Accounting for treatment changes

Both RCTs and real-world studies need to address treatment changes (e.g., non-adherence or switching) in the analysis. In most cases, an intent-to-treat (ITT) analysis is conducted, which assigns all outcomes to the treatment arm in which the patient was initially placed; this is sometimes interpreted as a conservative estimate of the treatment effect. It is possible to incorporate treatment changes and non-adherence during follow-up through an as-treated analysis, for example by censoring patients at the time of treatment change. Alternatively, as part of a more advanced approach, researchers can divide the follow-up period into time segments and measure the treatment and possible confounders separately in each segment. To better model real-world features such as medication management guidelines, researchers can introduce additional nuance by designing a per-protocol analysis with pre-specified treatment decision rules.

*The average treatment effect (ATE) is the difference in outcomes if every patient is given treatment A versus if every patient is given treatment B. One may also define the average treatment effect on the treated (ATT; the difference in outcomes if every patient who actually received treatment A had received treatment B instead) and its counterpart, the average treatment effect on the untreated (ATU). In an RCT with perfect covariate balance and treatment compliance, these effects are the same, but in observational data, the impact of selection bias and heterogeneous treatment effects across groups usually leads to divergence.

†An example of an estimand is: what is the average reduction in blood glucose over 1 year from initiation of drug class A among patients with existing type 2 diabetes, compared to patients initiating drug class B, in the period 2020 to 2023?

Steps 5 6 7

Creating a directed acyclic graph

Observational real-world studies are subject to multiple potential sources of bias, commonly grouped into confounding, selection, measurement, and time-related biases.[17] A practical first step in developing strategies to address threats to valid causal inference is to create a visual mapping of factors that may be related to exposure, outcome, or both (also called a directed acyclic graph or DAG).[18] DAGs typically draw upon prior literature, exploratory data analysis, and subject matter expertise to elucidate potential confounders of the exposure-outcome relationship. Confounders can be time-invariant or time-varying and observed or unobserved, and a DAG can help distinguish the importance of each for the research question.

Identifying and mitigating biases

Typical potential biases, such as confounding by time-invariant (baseline) characteristics, can be addressed through appropriate study design and statistical methods (e.g., the use of an active comparator, new user design combined with propensity score-based matching or weighting‡ to achieve balance in observed baseline confounders).[19,20] Researchers often use regression analysis to control for baseline confounding; propensity-score-based methods are preferred for this purpose, and, given the potential for other types of biases, regression typically does not, on its own, provide estimates suitable for causal inference.[19]

If the research question requires time-varying treatments (e.g., to account for treatment interruption or switching), propensity score-based covariate weights and censoring weights can be created based on the observed time-varying exposures and time-varying confounders and then incorporated into the analysis as part of a marginal structural model (MSM[21,22]).§ This approach allows researchers to address the bias of time-varying confounders and treatment-confounder feedback (e.g., when a medication impacts a biomarker that guides the selection of future treatment and the biomarker also affects the outcome of research interest).

Accounting for censoring

Patients in retrospective analyses are often censored due to limited follow-up associated with health plan disenrollment.** Such censoring can bias the analysis if it is related to patient characteristics (observable or unobservable) that are associated with the treatment and/or outcome. The effect of this dependent or 'informative' censoring can be mitigated by using censoring weights, which are estimated from the full population and incorporated into the outcomes analysis through inverse probability of censoring weighting. These weights can be applied to both ITT and as-treated analyses. In the case of as-treated analyses, they can also address bias due to informative censoring resulting from treatment discontinuation or switching.

‡ Weighting with propensity scores is referred to as inverse probability of treatment weighting (IPTW).

§ MSMs represent the most frequent model for handling time-varying exposures and confounders; others include g-computation and g-estimation (structural nested models). Note that MSMs are a class of models that can address time-invariant analyses as well (IPTW is a type of MSM).

**Alternatively, patients with limited follow-up may be excluded completely from the study, which can generate selection bias if the loss to follow-up is related to the exposure and/or outcome of the study.

Step 8

Sensitivity analysis

All research designs are built on several assumptions and involve trade-offs across methodologies. Sensitivity analysis is essential in understanding the uncertainty of study results (beyond those coming from statistical sampling uncertainty). Common examples of sensitivity analyses in comparative observational studies include modifying the study inclusion/exclusion criteria and using alternate methodologies to adjust for baseline confounding. Another type of sensitivity analysis seeks to quantify residual systematic bias.[23-25] Examples of this quantitative bias analysis include the assessment of unmeasured confounding and the incorporation of validation study results in adjusting for variable misclassification.[26-29] Online tools are increasingly available for this purpose, and the results can help to inform the research audience about potential study limitations stemming from residual systematic bias.

Carelon Research's expertise

Carelon Research has extensive experience applying causal study designs to real-world observational healthcare data. This section presents three case studies from our work with life science companies and payers to illustrate the application of advanced causal study designs and the generation of high-quality insights.

Case study #1

Research question

Does a new hormonal drug cause a higher risk of endometrial or breast cancer compared to existing hormone replacement therapies?

Causal inference challenges

Confounding by indication, informative censoring, prevalent-user bias, outcome misclassification

Solutions

Carelon Research researchers utilized multiple real-world databases and employed an active comparator, new-user design with propensity score matching to balance patient characteristics at the time of treatment initiation. In addition to the main analysis, the team conducted a variety of sensitivity and bias analyses to test the implicit assumptions of the study design. For example, to address the potential for overcounting cases due to erroneous or tentative ('rule out') diagnoses in claims, a claims-based algorithm was developed to identify endometrial cancer cases and validated with a review of patients' medical records. Using the parameters estimated in the validation study, a quantitative bias analysis was applied to the main study findings to examine the potential for outcome misclassification bias.

Results

In the main analysis, the study team found a slightly higher rate of endometrial cancer but a lower rate of breast cancer among users of the new hormonal drug relative to existing hormone replacement therapies. Quantitative bias analyses demonstrated that the observed effects were unlikely to be influenced by outcome misclassification. This research was presented at the 2021 International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE), and a manuscript has been submitted.

Case study #2

Research question

Does adherence to a maintenance medication regimen among patients with chronic obstructive pulmonary disease (COPD) result in clinical and economic benefits to support the launch of clinical interventions to improve adherence?

Causal inference challenges

Time-varying exposure and confounding, informative censoring

Solutions

Carelon Research researchers selected a population composed of COPD patients aged ≥40 years on a medication maintenance regimen and examined their medication adherence as a time-varying exposure on a daily rolling basis to delineate adherent vs. non-adherent follow-up time periods. MSMs were employed to examine the causal impact of adherence on outcomes in the presence of time-invariant and time-varying confounders by using repeated propensity-score-based weighting. This novel methodology applied an as-treated approach to account for the dynamic nature of adherence and its interaction with other time-varying patient factors that affect adherence and health and cost outcomes. This design also addressed concerns related to survival bias and temporality, which are commonly observed in prior literature examining the impact of adherence.

Results

The study team found that adherence to a COPD maintenance regimen resulted in meaningful clinical and economic benefits compared to non-adherence. This study laid the analytic groundwork for robustly assessing the effect of medication adherence on outcomes and is transferrable to different therapeutic areas. This research was presented at the 2023 Academy of Managed Care Pharmacy Annual Meeting, and a manuscript is under development.

Case study #3

Research question

Does a national insurer's Pay-for-Performance (P4P) program for oncology affect the prescribing of evidence-based cancer drugs and cancer care spending?

Causal inference challenges

Measured and unmeasured confounding, outcome misclassification, selection bias

Solutions

Carelon Research researchers used various methods, including a difference-in-differences design and event study analysis, to account for confounding based on patient and provider characteristics and self-selection of physicians who chose to participate in the P4P program. The analysis leveraged the geographically staggered, time-varying rollout of the P4P program. Potential misclassification of oncolytic regimens in claims was assessed through comparison with gold-standard registry data collected through the P4P program.

Results

The P4P program was associated with an increase in evidence-based regimen prescribing as well as in cancer drug spending and patient out-of-pocket spending, but no changes in total healthcare spending. The claims-based algorithm to identify oncolytic regimens had high concordance with the registry data. This research was presented at the 2020 Annual Meeting of the American Society of Clinical Oncology and published in the Journal of Clinical Oncology.[30]

Expertise tailored to you

Carelon Research's scientists curate the most appropriate research questions and identify the optimal study design and analytic approach to address specific research needs — from satisfying regulatory requirements to generating evidence to support important public health concerns.

Contact us at
rwe@carelon.com

Additional resources on causal inference

- Carelon Research ISPOR workshop presentation: Best practices for causal study designs using real-world data https://www.ispor.org/docs/default-source/intl2022/healthcore-umb-us-ispor-2022-causal-workshop-29apr2022.pdf?sfvrsn=4dfd182d_0
- Carelon Research perspective's article: Causal inference for real-world evidence https://www.carelonresearch.com/perspectives/causal-inference-for-real-world-evidence

Six suggested articles for further reading

- Hernán MA. The C-word: Scientific euphemisms do not improve causal inference from observational data. *Am J Public Health.* 2018;108(5):616-619.
- Franklin JM, Platt R, Dreyer NA, London AJ, Simon GE, Watanabe JH, Horberg N, Hernandez A, Califf RM. When can nonrandomized studies support valid inference regarding effectiveness or safety of new medical treatments? *Clin Pharmacol Ther.* 2022;111(1):108-115.
- Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol.* 2019;19(1):53. Published 2019 Mar 11.
- Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. *BMJ.* 2017;359:j4587.
- Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. *Pharmacoepidemiol Drug Saf.* 2015;24(10):1009-1016.
- Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol.* 2014;43(6):1969-1985.

References

1. Burcu M, Dreyer NA, Franklin JM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. *Pharmacoepidemiol Drug Saf.* 2020;29(10):1228–1235.
2. Franklin JM, Pawar A, Martin D, et al. Nonrandomized real-world evidence to support regulatory decision making: Process for a randomized trial replication project. *Clin Pharmacol Ther.* 2020;107(4):817–826.
3. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin Pharmacol Ther.* 2017;102(6):924–933.
4. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health.* 2018;108(5):616–619.
5. Bakker E, Plueschke K, Jonker CJ, Kurz X, Starokozhko V, Mol PGM. Contribution of Real-World Evidence in European Medicines Agency's Regulatory Decision Making. *Clin Pharmacol Ther.* 2022.
6. Capkun G, Corry S, Dowling O, et al. Can we use existing guidance to support the development of robust real-world evidence for health technology assessment/payer decision-making? *Int J Technol Assess Health Care.* 2022;38(1):e79.
7. Concato J, Corrigan-Curay J. Real-world evidence — where are we now? *N Engl J Med.* 2022;386(18):1680–1682.
8. Franklin JM, Platt R, Dreyer NA, et al. When can nonrandomized studies support valid inference regarding effectiveness or safety of new medical treatments? *Clin Pharmacol Ther.* 2022;111(1):108–115.
9. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol.* 2016;183(8):758–764.
10. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol.* 2016;79:70–75.
11. Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. *BMJ.* 2022;378:e071108.
12. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ.* 2016;355:i4919.
13. Administration UFaD. E9(R1) Statistical principles for clinical trials: Addendum: Estimands and sensitivity analysis in clinical trials; International Council for Harmonisation; Guidance for Industry; Availability. In: Food and Drug Administration H, ed. 86 FR 26047. Vol Docket No. FDA-2017-D-6113. Washington, DC2021:26047-26048.
14. Lawrance R, Degtyarev E, Griffiths P, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *J Patient-Rep Outcomes.* 2020;4(1):68.
15. Little RJ, Lewis RJ. Estimands, estimators, and estimates. *JAMA.* 2021;326(10):967–968.
16. Turchin A, Kaul A. BESTMED: oBservational Evaluation of Second line Therapy MEdications in Diabetes. https://bestmed.org/. Published 2021. Accessed Nov 18, 2022.
17. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol.* 2019;19(1):53.
18. Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. *JAMA.* 2022;327(11):1083–1084.
19. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res.* 2011;46(3):399–424.
20. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med.* 2015;34(28):3661–3679.
21. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
22. Shinozaki T, Suzuki E. Understanding Marginal structural models for time-varying exposures: Pitfalls and tips. *J Epidemiol.* 2020;30(9):377–389.
23. Matsouaka RA, Atem FD. Regression with a right-censored predictor using inverse probability weighting methods. *Stat Med.* 2020;39(27):4001–4015.
24. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics.* 2000;56(3):779–788.
25. Matsuyama Y, Yamaguchi T. Estimation of the marginal survival time in the presence of dependent competing risks using inverse probability of censoring weighted (IPCW) methods. *Pharm Stat.* 2008;7(3):202–214.
26. Beachler DC, de Luise C, Jamal-Allial A, et al. Real-world safety of palbociclib in breast cancer patients in the United States: a new user cohort study. *BMC Cancer.* 2021;21(1):97.
27. Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. *Am J Epidemiol.* 1993;138(11):1007–1015.
28. Lash TL, Ahern TP, Collin LJ, Fox MP, MacLehose RF. Bias analysis gone bad. *Am J Epidemiol.* 2021;190(8):1604–1612.
29. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol.* 2014;43(6):1969–1985.
30. Bekelman JE, Gupta A, Fishman E, et al. Association between a national insurer's pay-for-performance program for oncology and changes in prescribing of evidence-based cancer drugs and spending. *J Clin Oncol.* 2020;38(34):4055–4063.