Published in final edited form as:

Med Care. 2016 April ; 54(4): e23–e29. doi:10.1097/MLR.000000000000011.

Distinguishing selection bias and confounding bias in comparative effectiveness research

Sebastien Haneuse, PhD

Department of Biostatistics, Harvard School of Public Health.

Abstract

Comparative effectiveness research (CER) aims to provide patients and physicians with evidence-based guidance on treatment decisions. As researchers conduct CER they face myriad challenges. While inadequate control of confounding is the most-often cited source of potential bias, selection bias which arises when patients are differentially excluded from analyses is a distinct phenomenon with distinct consequences: confounding bias compromises internal validity while selection bias compromises external validity. Despite this distinction, however, the label “treatment-selection bias” is being used in the CER literature to denote the phenomenon of confounding bias. Motivated by an on-going study of treatment choice for depression on weight change over time, we formally distinguish confounding and selection bias in CER. By formally distinguishing selection and confounding bias we clarify important scientific, design and analysis issues relevant to ensuring validity. First is that the two types of bias may arise simultaneously in any given study; even if confounding bias is completely controlled, a study may nevertheless suffer from selection bias so that the results are not generalizable to the patient population of interest. Second is that statistical methods used to mitigate the two biases are themselves distinct; methods developed to control one type of bias should not be expected to address the other. Finally, the control of selection and confounding bias will often require distinct covariate information. Consequently, as researchers plan future studies of comparative effectiveness, care must be taken to ensure that all data elements relevant to both confounding and selection bias are collected.

Keywords

Confounding bias; selection bias; comparative effectiveness research

INTRODUCTION

Comparative effectiveness research (CER) aims to provide patients and physicians with evidence-based guidance as they make treatment decisions^{1,2}. With the passing of the American Recovery and Reinvestment Act of 2009 and the Patient Protection and Affordable Care Act of 2010, and the recent creation of the Patient-Centered Outcomes Research Institute³, CER is a current national priority.

As researchers conduct CER, they face myriad challenges. Chief among these, arguably, is confounding bias which arises when factors that simultaneously affect treatment choice and the outcome are not adequately controlled. This lack of adequate control compromises internal validity, specifically whether or not the observed results reflect causation⁴. A second, distinct challenge is selection bias which arises when the observed patients are not representative, in some way, of the broader patient population of interest⁵. This lack of representativeness compromises external validity, specifically whether or not results based on a sub-sample of patients are generalizable. Despite the two phenomena being distinct, the labels “confounding bias” and “selection bias” are not always rigorously or appropriately employed. The recent Institutes of Medicine (IOM) report on CER, for example, commented: “... *the decision to rely on data from observational studies must be weighed against the possibility of misleading results. The main form of bias (selection bias) occurs when the factors causing a person to experience the intervention are associated with the patient’s prognosis*”⁶. Furthermore, the label “treatment-selection bias” is increasingly being used for confounding bias in the emerging comparative effectiveness literature^{7–15}. While this label makes explicit that confounding arises when one fails to adjust for certain factors that affect treatment selection, it increases the potential for confusion between confounding bias and selection bias. Without a clear distinction between the two phenomena, however, CER studies run the risk of (i) inadvertently ignoring selection bias and/or (ii) erroneously believing that statistical methods developed for confounding can be used to adjust for selection bias, and/or (iii) finding that information needed to adequately control for selection bias has not been collected.

Towards mitigating these risks, this paper provides a formal distinction between confounding and selection bias in CER. It also discusses a number of important related issues. Specifically, formally distinguishing confounding and selection bias helps emphasize the distinction across statistical methods used to control the two types of bias. This, in turn, has important consequences for study design and the collection of information on patients. Specifically, to adequately control confounding bias, data collection must include all factors that are related to both treatment choice and the outcome of interest; to adequately control selection bias, data collection must include all factors related to why certain patients participate in the study and others not. The work in this paper was motivated by on-going CER study of treatment for depression on weight change over time which we briefly introduce..

CER FOR DEPRESSION TREATMENT

Depression and obesity are major public health concerns^{16,17}. While the processes underlying their impact on health outcomes are the subjects of much recent research^{18–22}, there is growing evidence that the choice of antidepressant drug therapy influences changes in weight over time^{23,24}. With climbing rates of obesity¹⁷ and antidepressant agents the most commonly prescribed drugs in the US²⁵, understanding the impact of antidepressant choice on weight change over time is crucial to helping patients and physicians make informed decisions.

Study setting

We obtained funding from the National Institute of Mental Health to examine this question using data from Group Health, an integrated insurance and health care delivery system serving approximately 650,000 members in western Washington State. As part of its clinical systems, Group Health maintains numerous electronic databases including an electronic health record (EHR) based on EpicCare (Epic Systems Corporation of Madison, WI) and an electronic pharmacy database, with complete prescription information since 1993. Group Health also maintains administrative databases that track demographic data, inpatient treatment and outpatient encounter claims, insurance and enrollment status, and visit appointments.

Data abstraction

To study the relationship between antidepressant drug therapy and weight change, we considered adults aged 18–65 years with a diagnosis of depressive disorder and who initiated a new monotherapy episode of antidepressant drug treatment. A new monotherapy episode was defined as a dispensing for a single medication, without any other antidepressant medication dispensing in the prior 9-months. Restricting to new monotherapy episodes initiated between 01/2006–11/2009, we identified N=10,606 eligible patients in the Group Health EHR. Data was subsequently abstracted on covariates relevant to the goals of the study. In particular, as the primary outcome of interest is weight change between treatment initiation (baseline) and two years later, all weight measurements in the EHR over a two-year follow-up period were abstracted.

The potential for confounding and selection bias

Since data in the Group Health EHR arises during the course of clinical care, and treatment choices are made within this context, potential confounding is clearly an important consideration. Towards eventually addressing confounding bias, data on all covariates thought to be related to disease severity, co-morbid conditions, treatment choice, and weight change was abstracted for all N=10,606 patients, as well as information on their primary care provider at the time of treatment initiation.

Ideally, all N=10,606 patients would have complete data in the EHR on all relevant covariates. Perhaps not surprisingly, this was not the case. Crucially, there was substantial missing data on weight: only n=1,637 patients had both baseline and two-year weight measurements. That complete weight information is missing for n=8,969 of the N=10,606 patients identified via the inclusion/exclusion criteria questions the representativeness of the sub-sample and suggests the strong potential for selection bias.

TERMINOLOGY

Informally, addressing confounding bias and selection bias requires answering two key questions. For confounding bias, the relevant question is: *why did a patient receive one particular drug over any other?*; for selection bias, the relevant question is: *why do some patients have complete data and others not?* Towards formalizing these questions, we introduce terminology used throughout this manuscript. As we define and elaborate upon the

terminology, Figure 1 provides an overview of their definitions and the interplay between the various concepts.

The study population

A central premise of this paper is the existence of some well-defined patient population of interest, to whom the results are intended to be generalized. Typically, this population is defined via pre-specified scientific inclusion/exclusion criteria including having been diagnosed with specific disease conditions (e.g. via ICD-9 coding) and demographic characteristics (e.g. age, gender, race). Additionally, researchers often apply further practical inclusion/exclusion criteria. For example, researchers may identify and recruit study participants within some geographic region, from among those who participated in some parent research study (e.g. the Nurses Health Study), from among the members of a health plan (e.g. an HMO or the VA), or by enrollment within the Medicare system.

Whichever scientific and practical inclusion/exclusion criteria are used, we refer to the resulting patient population as the *study population*. In the antidepressants study, the study population consists of all adults aged 18–65 years with a diagnosis of depression at Group Health and who initiated a new episode of drug monotherapy between 01/2006–11/2009. Based on these criteria, as mentioned above, the resulting the study population consisted of N=10,616 patients.

The study sub-sample

Once the study population is identified, a typical CER study proceeds by selecting patients to be included/recruited and on whom complete data will be collected/abstracted. We refer to these patients as having been selected into the *study sub-sample*. In practice, the study sub-sample may be a random sample from the study population, proactively invited and recruited into the CER study. For CER studies based on an existing, parent research study or on an EHR, proactive patient recruitment may not be necessary and the study sub-sample could correspond to those patients with complete information on all relevant covariates. In the antidepressants study the n=1,637 patients with complete weight data (assuming they have complete information on all other relevant covariates) constitute the study sub-sample.

Treatment assignment and selection mechanisms

Finally, towards formally distinguishing confounding and selection bias we consider two key mechanisms. The first is the *treatment assignment mechanism*, which characterizes how patient-, physician- and system-level characteristics influence the decision-making process regarding which treatment any given patient is assigned. The second is the *selection mechanism*, which characterizes how patient-, physician-, system-level characteristics influence the decision-making process of whether or not a patient in the study population is selected into the study sub-sample.

CONFOUNDING AND SELECTION BIAS

Consider a comparative effectiveness study of the association between some treatment choice, denoted Rx, and an outcome of interest, denoted Y. In the antidepressants study, Rx

is the choice of antidepressant and Y is the two-year change in weight post-treatment initiation. Assuming a well-defined study population, to formally identify which patients are selected into the study sub-sample, let $S=0/1$ be a binary indicator of observance or selection: individuals selected and whose information is observed have $S=1$; individuals not selected and on whom (at least) some information is incomplete have $S=0$. In the following we use a series of simple directed acyclic graphs (DAGs) to formalize confounding and selection bias^{26,27}, to emphasize their distinction from each other and to illustrate that the two sources of bias may arise independently and simultaneously.

Confounding bias

Suppose a randomized trial is conducted to investigate the association between Rx and Y . Additionally, suppose patients are prospectively recruited from the study population via random sampling and that information on a collection of pre-treatment factors associated with the outcome, denoted by L , is collected. Since treatment assignment is random in the trial, estimates based on the study sample can be viewed as representing the causal effect of Rx on Y . That is, randomization guarantees no confounding bias (on average, at least) and internal validity of the study results is ensured.

Figure 2(a) provides a directed acyclic graph (DAG) for the randomized trial. That no arrows lead into Rx indicates that treatment assignment is independent of all other factors (that is, it is random). In contrast, Figure 2(d) provides the DAG for an observational study in which treatment assignment depends on at least one component of L . Since treatment assignment is not random, confounding bias will arise if one fails to adjust for L in subsequent analysis and, without appropriate adjustment, internal validity of the study will be compromised.

Structurally, the potential for confounding in Figure 2(d) is indicated by the arrows that simultaneously emanate from L into both Rx and Y . This structure is also present in Figures 2(e) and 2(f). Consequently, despite differences with Figure 2(d) in other parts of their structure, Figures 2(e) and 2(f) also represent studies in which treatment assignment was not random and for which there is the potential for confounding bias.

Selection bias

Returning to Figure 2(a), the presence of " $S=1$ " in the DAG serves to indicate that the effect of Rx on Y is being investigated solely using information from the study sub-sample. Specifically, the box around " $S=1$ " emphasizes that the selection into the study sub-sample is being conditioned upon^{26,27}. That there are no arrows leading into the boxed " $S=1$ ", however, indicates that mechanism driving selection is independent of Rx , L and Y ; as mentioned above, patients were recruited into the trial via random sampling. Since selection is random, the study sub-sample can be viewed as being representative of study population and the presence of the boxed " $S=1$ " in the DAG does not impact compromise the results. That is, random sampling of patients into the study sub-sample guarantees no selection bias (on average, at least) and external generalizability of the study results to the broader patient population of interest is ensured.

Now suppose that the study sub-sample was not obtained via random sampling and that some non-random selection mechanism drove whether or not a patient from the study population is actually in the study sub-sample. For example, suppose that whether or not a patient decides to participate in the randomized trial, after having been invited, depends on some component of L . Patients with greater pre-treatment disease severity may, for example, be less likely to agree to participate. Figure 2(b) provides a DAG for this setting, with the arrow from L into the boxed “ $S=1$ ” representing the dependence of selection on L . As a second example, consider a randomized trial in which patients were initially chosen at random from the study population and all agreed to participate. However, over time some patients drop out prior to the end of follow-up and the decision to drop out is dependent on their initial treatment assignment as well as their response to treatment. A study participant may, for example, decide to drop out if they experience some treatment-specific adverse side effect and/or if they do not respond to treatment as hoped. Figure 2(c) provides a DAG for this setting, with the arrows from Rx and Y into the boxed “ $S=1$ ” representing the dependence of the selection mechanism on treatment choice and response.

Structurally, Figures 2(b) and 2(c) share the common feature of having at least one arrow leading into the boxed “ $S=1$ ”. While the non-random selection mechanism differs between the two studies (one is a result of differential participation and the other differential drop-out), the upshot is the same: the study sub-sample, on whom complete data is available, is not representative of the study population. In this sense, both studies are subject to potential selection bias and, without appropriate adjustment external generalizability of the study results will be compromised. Despite this, however, treatment assignment is independent of all other factors in both figures, as evidenced by the lack of an arrow leading into Rx . Consequently, even though randomization guarantees internal validity, there is nevertheless the potential for external generalizability to be compromised in both Figures 2(b) and 2(c).

Simultaneous confounding and selection bias

Finally, consider Figures 2(e) and 2(f). As mentioned, both DAGs exhibit the potential for confounding bias due to the arrows simultaneously emanating from L into both Rx and Y . In contrast to Figure 2(d), however, the two DAGs also indicate the potential for non-random selection into the study sub-sample. Figure 2(e) represents an observational study with differential participation while Figure 2(f) represents an observational study with differential drop out. Consequently, Figures 2(e) and 2(f) represent different studies that are simultaneously subject to confounding bias and selection bias. That is, both internal validity and external generalizability are potentially compromised and would both need to be addressed in subsequent analyses.

STATISTICAL ANALYSES

While important conceptually, the distinction between confounding and selection bias is also important from the perspective of statistical analyses. To help ground a discussion of these issues, Figure 3 provides a simplified DAG for the motivating antidepressants study. From the Figure, the true data model is a linear regression for the mean two-year weight change is a function of antidepressant treatment choice and three covariates: baseline smoking, baseline weight and gender. In this simplified setting, treatment is binary (say antidepressant

A versus antidepressant B) and is assigned via some mechanism that depends on baseline smoking and weight but not gender. Furthermore, whether or not an individual patient has complete data in the EMR (i.e. both baseline and two-year weight measurements) is assumed to be dependent on treatment choice, baseline smoking status and gender. From the discussion in the previous section, the DAG in Figure exhibits structure that is consistent with the study potential being subject to both confounding and selection bias.

Confounding bias

For the purpose of statistical control of confounding, there is a long history of methodologic research including recent work that is specific to CER^{28,29}. For the most part, methods fall into one of two general approaches: regression adjustment and propensity score analyses^{30,31}. Regression adjustment relies on building a model for the relationship between the treatment and outcome, and including potential confounders in the specification of that model. For the DAG in Figure 3, both the “True data model” and the “Regression adjustment model” would suffice for this strategy since (at a minimum) they both include the two confounders; since gender is not a confounder, its inclusion in a model for the outcome is unnecessary from the perspective of confounding control although it will increase statistical efficiency. While regression adjustment relies on a model for the outcome, propensity score analyses rely on a model for the probability of treatment as a function of potential confounders. In Figure 3, treatment assignment depends solely on baseline smoking and weight; for a binary treatment choice, a logistic regression model such as the one labeled “Propensity model” may be used. Fitted values from this model (i.e. predicted probabilities of treatment for any given patient) are referred to as propensity scores, which can then be used in various ways to control confounding bias, including stratification and inverse-probability weighting.

In practice, the regression adjustment and propensity score analyses perform equally well in most settings³⁰. Crucial to both approaches, however, is that the treatment assignment mechanism is understood. Without this understanding, one would not know which covariates to include in either of the outcome or propensity score models.

Selection bias

In contrast to confounding bias, methods specific to selection bias are relatively sparse, particularly in the CER literature. Nevertheless, a useful strategy for the statistical adjustment of selection bias is to view the patients in the study population who were not selected as having missing data. When cast as a missing data problem, selection bias can then be addressed using the broad range of methods developed for more general missing data settings. One such method is multiple imputation^{31,32}. Unfortunately, however, its application in the selection bias context will often be challenging because it involves imputing all information for all patients in the study population who were not selected into the study sub-sample. A second missing data approach is inverse-probability weighting in which patients who are observed in the study sub-sample are re-weighted in an effort to “reconstruct” the original study population^{32,33}. The weights are taken from the fit of a model that treats whether or not a patient is selected into the sub-sample as the outcome. For the DAG in Figure 3, this corresponds to fitting a model for the probability of having

complete data in the EMR; the logistic regression labeled “Selection model” provides one possible choice that would ensure that all factors relevant to selection are considered.

Data considerations

Beyond the control of confounding bias and selection bias requiring the specification of distinct mechanisms/models, practically, statistical analyses also require fundamentally different data. Specifically, the control of confounding bias requires data on all covariates relevant to the treatment assignment mechanism, on all patients in the sub-sample. That is, the “Regression adjustment model” and/or the “Propensity model” in Figure 3 would only be fit to those patients with complete data in the EMR (i.e. those with $S=1$). In contrast, the control of selection bias requires data on all covariates relevant to selection mechanism, on all patients in the study population. That is, the “Selection model” in Figure 3 would be fit to all patients in the EMR (i.e. those with $S=0$ and $S=1$).

STUDY DESIGN

That the analysis approaches used to control confounding bias and selection bias require different data has important consequences for study design. As comparative effectiveness studies are developed, submitted for review and funding, considerable attention is typically given to ensuring that all confounders are identified and included in the data collection plan. In the event that data on an important confounder is missing, the study results will suffer from unmeasured confounding bias. In contrast, less emphasis is typically placed on ensuring that all factors involved in the selection process are collected. This may be due, in part, to the difficult challenge of needing data on these factors for all patients in the study population. In some settings, this information may be readily available. In the antidepressants study, for example, if the patients’ age, gender or concurrent co-morbid conditions determine, in part, whether or not a weight measurement is obtained at baseline, this information is available in the EHR. Similarly, if weight measurements are only collected during primary care visits, as opposed to specialty care visits, information in the EHR can be used to identify the visit type, which can then be included in a model for selection. In other settings, however, information relevant to selection may not be available in the EHR. Most problematic for the antidepressants study is that a patients’ weight itself may be a determinant of whether or not weight is missing in the EHR and, hence, whether or not that particular patient is included in the study sub-sample. To use the standard missing data nomenclature, this is an informative missingness or missing-not-at-random^{31,32}. Unfortunately, given data in the EHR alone, analyses that attempt to control selection bias would be inadequate and, analogous to unmeasured confounding, residual selection bias would manifest. Resolving this challenge requires additional data collection, beyond the information available in the EHR, specific to the control of selection bias. This is likely best achieved through explicit consideration of the selection mechanism and careful planning during the design phase of any given CER study. How to do this efficiently, however, is currently an open methodologic problem.

DISCUSSION

Researchers conducting CER must navigate a wide range of phenomena that result in bias and compromise validity. Unfortunately, the literature is not wholly consistent in the labels used to describe and distinguish these phenomena. The bias that results from inadequate adjustment of a covariate that is simultaneously predictive of treatment and outcome, for example, has been referred to as “confounding bias”, “confounding by indication bias” and “treatment-selection bias”. Furthermore, as Hernan et al²⁷ point out, each of the following phenomena/biases can be viewed as being structurally equivalent: inappropriate selection of controls in case-control studies, bias differential drop out, volunteer bias, healthy worker bias and non-response bias. That numerous labels are used for the same phenomenon runs the risk of certain biases being overlooked or inadequately handled as studies are designed, conducted, analyzed and interpreted. In particular, since the control of selection bias and the control of confounding bias use fundamentally different data and statistical analysis approaches, maintaining a formal distinction is essential. Furthermore, as CER matures new methods will likely need to be developed. Compared to the current trend in research for confounding bias, however, the development of methods for the control of selection bias in CER, including sensitivity analyses, has been sparse at best. Three recent reviews of methods for CER, for example, provided virtually no discussion of selection bias or external validity as a methodologic concern^{29,34,35}.

Confounding bias and selection bias are only two of the potential challenges that CER researchers face. Others include misclassification, measurement error and self-report bias, and range of issues specific to the use of large electronic health/administrative databases³⁶. With the current emphasis on patient-centered research and personalize medicine, treatment effect heterogeneity has also been identified as an important challenge in CER^{37–39}. In the literature it is common for studies to report the overall association between treatment choice and outcomes, the so-called main effect. While this can loosely be interpreted as the average effect across all patients in the population, if the treatment effect is truly heterogeneous the overall effect may not be relevant to any given patient. Consequently, treatment effect heterogeneity has implications for generalizability of study results although this is a distinct phenomenon to the compromised generalizability induced by selection bias. Indeed, external validity of study results as we have defined them (i.e. the ability to directly translate results from a sub-sample to the entire patient population) may be compromised regardless of whether or not treatment effects are heterogeneous.

CONCLUSION

Traditionally, confounding bias has been viewed as the greatest threat to validity of observational comparative effectiveness research. While this may indeed be the case, if study results are to be translatable into clinical practice and, ultimately, useful to patients, researchers must consider all sources of bias. Crucially, consideration needs to be given at the planning and design phase to ensure that researchers have access to all of the information needed to perform statistical adjustments.

Acknowledgments

FUNDING: This work was funded, in part, by grant R01 MH083671 from the National Institutes of Mental Health (NIHM; PI David Arterburn).

REFERENCES

1. Steinbrook R. Health care and the American Recovery and Reinvestment Act. *The New England Journal of Medicine*. 2009; 360(11):1057–1060. [PubMed: 19224738]
2. Federal Coordinating Council for Comparative Effectiveness Research (U.S.), United States. President., United States. Congress., United States. Dept. of Health and Human Services. Report to the President and the Congress. Washington, DC: US Dept. of Health and Human Services; 2009.
3. Patient-Centered Outcomes Research Institute. <http://www.pcori.org/>.
4. Rothman, K.; Greenland, S.; Lash, TL. *Modern Epidemiology*. N/A. 3 ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
5. Rothman, KJ.; Greenland, S.; Lash, TL. *Modern epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
6. Institute of Medicine (U.S.). Initial national priorities for comparative effectiveness research. Washington, DC: National Academies Press; 2009. Committee on Comparative Effectiveness Research Prioritization.
7. Novikov I, Kalter-Leibovici O. Analytic approaches to observational studies with treatment selection bias. *JAMA : the journal of the American Medical Association*. 2007; 297(19):2077. auhor reply 2078. [PubMed: 17507342]
8. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. 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 : the journal of the American Medical Association*. 2007; 297(3):278–285. [PubMed: 17227979]
9. Griswold ME, Localio AR, Mulrow C. Propensity score adjustment with multilevel data: setting your sites on decreasing selection bias. *Annals of internal medicine*. 2010; 152(6):393–395. [PubMed: 20231571]
10. Lalani T, Cabell CH, Benjamin DK, Lasca O, Naber C, Fowler VG Jr, Corey GR, Chu VH, Fenely M, Pachirat O, Tan RS, Watkin R, Ionac A, Moreno A, Mestres CA, Casabe J, Chipigina N, Eisen DP, Spelman D, Delahaye F, Peterson G, Olaison L, Wang A. Analysis of the impact of early surgery on in-hospital mortality of native valve endocarditis: use of propensity score and instrumental variable methods to adjust for treatment-selection bias. *Circulation*. 2010; 121(8): 1005–1013. [PubMed: 20159831]
11. Ounpraseuth S, Gauss CH, Bronstein J, Lowery C, Nugent R, Hall R. Evaluating the effect of hospital and insurance type on the risk of 1-year mortality of very low birth weight infants: controlling for selection bias. *Medical Care*. 2012; 50(4):353–360. [PubMed: 22422056]
12. Suh HS, Hay JW, Johnson KA, Doctor JN. Comparative effectiveness of statin plus fibrate combination therapy and statin monotherapy in patients with type 2 diabetes: use of propensity-score and instrumental variable methods to adjust for treatment-selection bias. *Pharmacoepidemiology and drug safety*. 2012; 21(5):470–484. [PubMed: 22461130]
13. Hernandez AF, Mi X, Hammill BG, Hammill SC, Heidenreich PA, Masoudi FA, Qualls LG, Peterson ED, Fonarow GC, Curtis LH. Associations between aldosterone antagonist therapy and risks of mortality and readmission among patients with heart failure and reduced ejection fraction. *JAMA : the journal of the American Medical Association*. 2012; 308(20):2097–2107. [PubMed: 23188026]
14. Weintraub WS, Grau-Sepulveda MV, Weiss JM, O'Brien SM, Peterson ED, Kolm P, Zhang Z, Klein LW, Shaw RE, McKay C, Ritzenthaler LL, Popma JJ, Messenger JC, Shahian DM, Grover FL, Mayer JE, Shewan CM, Garratt KN, Moussa ID, Dangas GD, Edwards FH. Comparative effectiveness of revascularization strategies. *The New England journal of medicine*. 2012; 366(16):1467–1476. [PubMed: 22452338]

15. Wang L, Wei W, Miao R, Xie L, Baser O. Real-world outcomes of US employees with type 2 diabetes mellitus treated with insulin glargine or neutral protamine Hagedorn insulin: a comparative retrospective database study. *BMJ open*. 2013; 3(4)
16. Stunkard AJ, Faith MS, Allison KC. Depression and obesity. *Biological Psychiatry*. 2003; 54(3): 330–337. [PubMed: 12893108]
17. Flegal KM, Carroll MD, Ogden CL, Curtin LR. Prevalence and trends in obesity among US adults, 1999–2008. *Journal of the American Medical Association*. 2010; 303(3):235–241. [PubMed: 20071471]
18. Onyike CU, Crum RM, Lee HB, Lyketsos CG, Eaton WW. Is obesity associated with major depression? Results from the Third National Health and Nutrition Examination Survey. *American Journal of Epidemiology*. 2003; 158(12):1139–1147. [PubMed: 14652298]
19. Zhao G, Ford ES, Li C, Tsai J, Dhingra S, Balluz LS. Waist circumference, abdominal obesity, and depression among overweight and obese U.S. adults: National Health and Nutrition Examination Survey 2005–2006. *BMC Psychiatry*. 2011; 11:130. [PubMed: 21834955]
20. Vogelzangs N, Kritchevsky SB, Beekman AT, Brenes GA, Newman AB, Satterfield S, Yaffe K, Harris TB, Penninx BW. Obesity and onset of significant depressive symptoms: results from a prospective community-based cohort study of older men and women. *The Journal of Clinical Psychiatry*. 2010; 71(4):391–399. [PubMed: 20021992]
21. Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers P, Penninx BW, Zitman FG. Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. *Archives of general Psychiatry*. 2010; 67(3):220–229. [PubMed: 20194822]
22. Faith MS, Butryn M, Wadden TA, Fabricatore A, Nguyen AM, Heymsfield SB. Evidence for prospective associations among depression and obesity in population-based studies. *Obesity reviews : an official journal of the International Association for the Study of Obesity*. 2011; 12(5):e438–e453. [PubMed: 21414128]
23. Serretti A, Mandelli L. Antidepressants and body weight: a comprehensive review and meta-analysis. *The Journal of Clinical Psychiatry*. 2010; 71(10):1259–1272. [PubMed: 21062615]
24. Patten SB, Williams JV, Lavorato DH, Khaled S, Bulloch AG. Weight gain in relation to major depression and antidepressant medication use. *Journal of Affective Disorders*. 2011; 134(1–3): 288–293. [PubMed: 21774992]
25. Paulose-Ram R, Safran MA, Jonas BS, Gu Q, Orwig D. Trends in psychotropic medication use among U.S. adults. *Pharmacoepidemiology and Drug Safety*. 2007; 16(5):560–570. [PubMed: 17286304]
26. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999; 10(1):37–48. [PubMed: 9888278]
27. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004; 15(5):615–625. [PubMed: 15308962]
28. Schneeweiss S. Developments in post-marketing comparative effectiveness research. *Clinical Pharmacology and Therapeutics*. 2007; 82(2):143–156. [PubMed: 17554243]
29. Sox HC, Goodman SN. The methods of comparative effectiveness research. *Annual Review of Public Health*. 2012; 33:425–445.
30. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007; 26(4):734–753. [PubMed: 16708349]
31. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996; 91(434):473–489.
32. Little, R.; Rubin, D. *Statistical analysis with missing data*. N/A. 2 ed. Hoboken, NJ: John Wiley and Sons; 2002.
33. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89(427):846–866.
34. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. *Journal of Clinical Oncology*. 2012; 30(34):4215–4222. [PubMed: 23071228]
35. Armstrong K. Methods in comparative effectiveness research. *Journal of Clinical Oncology*. 2012; 30(34):4208–4214. [PubMed: 23071240]

36. Hernan MA. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. *Epidemiology*. 2011; 22(3):290–291. [PubMed: 21464646]
37. Garber AM, Tunis SR. Does comparative-effectiveness research threaten personalized medicine? *The New England Journal of Medicine*. 2009; 360(19):1925–1927. [PubMed: 19420360]
38. Kaplan SH, Billimek J, Sorkin DH, Ngo-Metzger Q, Greenfield S. Who can respond to treatment? Identifying patient characteristics related to heterogeneity of treatment effects. *Medical Care*. 2010; 48(6 Suppl):S9–S16. [PubMed: 20473205]
39. Willke RJ, Zheng Z, Subedi P, Althin R, Mullins CD. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Medical Research Methodology*. 2012; 12:185. [PubMed: 23234603]

Figure 1.

Interplay between the study population and study sub-sample in the context of distinguishing external from internal validity. Numbers in the parentheses correspond to the observed sample sizes in the motivating comparative effectiveness study of treatment for depression and two-year weight change.

(a) RT with random sampling

Potential confounding bias: No
Potential selection bias: No

(b) RT with differential participation

Potential confounding bias: No
Potential selection bias: Yes

(c) RT with differential dropout

Potential confounding bias: No
Potential selection bias: Yes

(d) OS with random sampling

Potential confounding bias: Yes
Potential selection bias: No

(e) OS with differential participation

Potential confounding bias: Yes
Potential selection bias: Yes

(f) OS with differential dropout

Potential confounding bias: Yes
Potential selection bias: Yes

Figure 2.

Directed acyclic graphs illustrating the potential for confounding bias and selection bias under a randomized trial (RT) or an observational study (OS), with various scenarios for the selection mechanism. In each sub-figure, Rx is treatment choice, Y is the outcome of interest and L is a collection of factors related to the outcome and, possibly, treatment choice. The boxed "S=1" indicates that analyses are only performed on patients selected into the study sub-sample.

True data model:

$$E[Y] = \beta_0 + \beta_x Rx + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3$$

Regression adjustment model:

$$E[Y] = \beta_0 + \beta_x Rx + \beta_1 L_1 + \beta_2 L_2$$

Propensity model:

$$\text{logit } P(Rx=1) = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$$

Selection model:

$$\text{logit } P(S=1) = \gamma_0 + \gamma_x Rx + \gamma_1 L_1 + \gamma_3 L_3$$

Figure 3.

A simplified directed acyclic graph for the motivating comparative effectiveness study of treatment for depression (Rx) on two-year weight change (Y). Baseline smoking (L₁) and weight (L₂) are confounders of the association of interest; gender (L₃) is associated with weight change but independent of treatment choice. Baseline smoking and gender are determinants of selection into the study sub-sample but being associated with whether or not

a patient has complete data in the EMR. Also shown are various models relevant to the adjustment of confounding bias and selection bias.