Chapter 5. Comparator Selection

Soko Setoguchi, M.D., Dr.P.H.
Duke Clinical Research Institute, Durham, NC

Tobias Gerhard, Ph.D.
Rutgers University, New Brunswick, NJ

Abstract

This chapter discusses considerations for comparator selection in observational comparative effectiveness research (CER). Comparison groups should reflect clinically meaningful choices in real world practice and be chosen based on the study question being addressed. Recognizing the implications and potential biases associated with comparator selection is necessary to ensure validity of study results; confounding by indication or severity and selection bias (e.g., healthy user bias) is particularly challenging, especially with comparators of different treatment modalities. Confounding by indication can be minimized by choosing a comparator that has the same indication, similar contraindications, and a similar treatment modality (when possible). In fact, comparing a treatment with a clinically meaningful alternative treatment within the same or a similar indication is the most common scenario in CER, and also typically the least biased possible comparison.When carefully planned, comparisons of different treatment types are possible with adequate study design, execution, and appropriate analytic methods. However, we note that certain comparisons or study questions may not be feasible or valid to be answered in observational CER studies due to potentially uncontrollable bias. Other aspects to consider when choosing a comparator include clearly defining the indication, initiation period, and exposure window for each group. The appropriate dose/intensity of each exposure should be as comparable as possible and nonadherence should be considered (although not necessarily adjusted for). This chapter concludes with guidance and key considerations for choosing a comparison group for an observational CER protocol or proposal.

Introduction

In comparative effectiveness research (CER), the choice of comparator directly affects the validity of study results, clinical interpretations, and implications. When formulating a research question, therefore, careful attention to proper comparator selection is necessary.

Treatment decisions are based on numerous factors associated with the underlying disease and its severity, general health status or frailty, quality of life, and patient preferences—a situation that leads to the potential for confounding by indication or severity and selection bias. Recognizing the implications and potential biases associated with comparator selection is critical for ensuring the internal validity of observational CER studies. The first section of this chapter, "Choosing the Comparison Group in CER," begins by describing these biases, and discusses the potential for bias associated with different comparison groups (e.g., no intervention, usual care, historical controls, and comparison groups from other data sources).

Defining the appropriate dose, intensity of treatment, and exposure window for each comparator group is also critical for ensuring the validity of observational CER.The second section of this chapter, "Operationalizing the Comparison Group in CER," discusses these considerations for operationalizing comparison groups, and concludes with special considerations that apply to CER studies comparing different treatment modalities.

Choosing the Comparison Group in CER

Link to Study Question

In CER, comparison groups should reflect clinically meaningful choices in real world practice. The

selection of comparison group(s) is thus directly linked to the study question being addressed. Importantly, some comparisons or study questions may not be feasible or valid to be answered in observational CER studies due to expected intractable bias or confounding.

Consequences of Comparator Choice

Confounding

Confounding arises when a risk factor for the study outcome of interest (benefit or harm) directly or indirectly affects exposure (e.g., treatment assignment). Because clinicians routinely make treatment decisions based on numerous factors associated with the underlying disease and its severity, confounding by indication or severity poses a significant threat to the validity of observational CER (see chapter 2 for a detailed discussion). It is therefore vital to appreciate the relationship between confounding and comparator choice. The existence and magnitude of confounding for any given pair of treatments and outcome is directly affected by the choice of the comparator. For example, when comparing the adverse metabolic consequences of individual antipsychotic medications in patients with schizophrenia or bipolar disorder, body mass index (BMI) is an important potential confounder because it is a strong and established risk factor for adverse metabolic outcomes such as type 2 diabetes and plausibly affects the choice of agent. However, the expected magnitude of confounding by BMI strongly depends on the specific drugs under study. A comparison between aripiprazole, an antipsychotic agent with a relatively favorable metabolic safety profile, and olanzapine, an agent that exhibits substantial metabolic adverse effects, may be strongly confounded by BMI, as most clinicians will try to avoid olanzapine in patients with increased BMI. In contrast, a comparison between aripiprazole and another antipsychotic agent with less metabolic concerns than olanzapine, such as ziprasidone, may be subject to confounding by BMI but to a much lesser degree.

The magnitude of potential confounding generally is expected to be smaller when the comparator (1) has the same indication, (2) has similar contraindications, (3) shares the same treatment modality (e.g., tablet or capsule), and (4) has similar adverse effects. Therefore, selection of a comparator of the same treatment modality (e.g., drug vs. drug) and same class within the modality (e.g., β-blocker) may result in less confounding than comparison across different treatment modalities or drug classes in general. However, many exceptions exist (e.g., the antipsychotic example above), and assessments should be made individually for each treatment comparison of interest. To understand the potential consequences of comparator choices on confounding, a thorough understanding of clinical practice, data sources, and methods is necessary. If suspected confounders are available in the data, investigators can empirically evaluate the extent that the distribution of these confounders differs between the exposure of interest and the comparator(s).

Propensity score distribution plots by exposure status are particularly useful in this context because they allow simple evaluation of the joint differences of many potential confounders between treatments. Areas of nonoverlap between the propensity score distribution in the treatment and comparator group identify individuals who, based on their baseline characteristics, would either always or never be exposed to the treatment under study, and thus cannot be compared without potential for significant bias.[1] If potential confounders are not available in the data, practical clinical insight and qualitative health services research should be used to form an impression of the expected magnitude of confounding for a given treatment comparator pair. Sensitivity analyses should then be used to quantify the effects of such unmeasured confounding under different sets of assumptions. (See chapter 11 for further discussion).[2]

While a thorough understanding of the impact of comparator choice on the expected magnitude of confounding is critical, the comparator choice should be primarily driven by a comparative effectiveness question that has been prioritized by the informational need of the stakeholder community. We do not advocate for minimizing confounding through a comparator choice that might change the original study question. A critical assessment of the expected magnitude of confounding for the comparison group of choice, however, should guide decisions of study design, particularly (1) the need to obtain additional

covariate information if confounding is judged to be uncontrollable in the available data (despite use of advanced analytic methods, such as propensity scores and other approaches described further in chapter 10); and (2) the need for randomization if confounding is judged uncontrollable in any observational study design even with additional data collection (despite use of advanced analytic methods).

Misclassification

Misclassification is one of the major threats to validity in observational CER studies and is discussed in more detail in chapter 4 and chapter 6. In the context of selecting comparison groups for CER, it is important to appreciate that exposure misclassification is often not binary but rather more complex, as each group (exposure and comparison group) typically represents an active treatment, and as nonuse of the exposure treatment does not imply use of the comparator treatment. For example, consider an epidemiologic study of the effect of treatment A (exposed) on outcome Y. If nonexposure to A is the comparison of interest, this category of exposure is directly dependent on exposure to A, as each subject is either exposed or unexposed to A. Therefore, misclassification of exposure A would affect the number of those identified as having A (exposure group) *and* those without A (comparison group). However, in a CER comparing the effects of drug A versus drug B, misclassification of exposure A would not necessarily affect the number of patients with drug B (comparison), as exposure to A is largely independent of exposure to B.

In observational CER, the assessment of exposure misclassification has to be made for the exposure and comparison group independently, and it is important to recognize that the degree of misclassification can be different in the two groups, especially when the comparison groups come from different treatment modalities (e.g., drug vs. device). Generally, the more similar the treatment under study and the comparator are in terms of treatment modality and dosage form, the less likely it is that exposure or comparator misclassification is different. For example, there is little reason to expect that the degree of exposure misclassification would substantially differ between the comparison groups in a claims-based study comparing two oral pharmacologic treatments, as information on drug exposure is equally retrieved from pharmacy billing claims for both groups. However, in a comparison between an oral medication for chronic diseases and a long-term injectable, the degree of misclassification may be significantly larger for patients treated with the oral dosage form mainly due to the different way of administering the drugs (patient vs. physician) and sources of information (drug dispensing records vs. office visit records).

Spectrum of Possible Comparisons

Comparison interventions may include medications, procedures, medical and assistive devices and technologies, behavioral change strategies, and delivery systems. Under certain circumstances, no intervention, usual care, historical controls, or comparison groups from other data sources may be appropriate and justified for comparative effectiveness questions. It is again important to recognize that comparator choice is directly linked to the comparative effectiveness question under study. In this section, we will discuss methodological considerations for the choice of different comparison groups.

Alternative Treatments

Comparison of a treatment with a clinically meaningful alternative treatment within the same or a similar indication is the most common scenario in CER and also typically the least biased comparison. Multiple modalities and options are often available to treat or diagnose the same condition or indication. Therefore, in many clinical circumstances, “no treatment” or “no testing” may not meet usual standards of care, and comparisons with alternative treatment options may be more clinically meaningful and methodologically valid. Comparison with alternative treatment or testing within the same or similar indication is usually a better choice from a methodologic standpoint than comparison with an untreated/not tested group, as confounding by indication may be nonexistent or at least reduced in the former comparison. However, when different treatments or testing modalities are recommended for patients with varying levels of severity of the underlying condition, comparisons within the same indication may still result in confounding by severity when not adequately controlled through design or analysis.

No Treatment

Comparison with no treatment or no testing may be appropriate in certain clinical situations. When a comparison with no treatment is a clinically appropriate question, researchers may define the no-treatment group as the absence of exposure or, alternatively, as the absence of exposure *and* use of an unrelated treatment (an active comparator) within the same source population. Active comparators are users of treatments that are not associated with indications for the exposure treatment and, importantly, have no effect on the outcome of interest (supported by available evidence).[3] The goal of employing active comparators who are likely to have similar characteristics with the exposure treatment users is to remove or minimize bias due to unobserved or incompletely observed differences between treated and untreated patients. For example, in a study assessing the risk of cancer in statin use,[4] users of glaucoma drugs (like statins, a preventive medication class less likely to be used in frail elderly patients[5]), were employed as an active comparison group with an aim to control for potential bias due to statin users' being more health-seeking and more adherent to screening procedures and other recommendations than nonusers.[3] While this approach is likely to have greater applicability to questions of safety than CER, it may warrant consideration in addressing some CER questions.

Another important consideration, when "no treatment" is appropriate as a comparison group, is how to select time zero for the no-treatment group. When an active comparison group is employed, the choice of time zero is naturally determined as the start of the active treatment. When a no-treatment comparison group is selected, one way to choose time zero is to identify the day a health care professional made a no-treatment decision. This way, both cohorts will have a meaningful inception date for the start of exposure status and outcome identification. However, in many clinical scenarios, such a date may not exist, as no treatment is often considered for patients in early stage of disease progression. Additionally, even if such a date exists, it may be difficult to identify in the available data. A second way to handle this is to allow a different time zero for the treatment and no-treatment groups (time-varying exposure status), and to carefully consider allocation of person-time to avoid immortal person-time bias.[6] In a third design strategy, it is possible to align the person-time and events appropriately by a choice of time scale in a Cox proportional hazard regression.[7] Researchers should realize that the choice of time zero in a no-treatment comparison group can induce bias, and careful considerations are needed to select clinically appropriate time zero and/or to avoid immortal person-time bias, as choice for no treatment is often related to disease stage and progression and therefore outcomes.

Usual or Standard Care

When a new treatment or testing modality becomes available, patients and health care providers may ask a question about the effectiveness of the new treatment when added to the usual or standard care. While this question is legitimate and important, operationalizing the question into an answerable research question requires a clear definition of "usual or standard care," including a valid operational definition of when usual care was initiated. The standard care could be no treatment or no testing, a single treatment or testing, or a set of existing treatment or testing modalities. In the real world, patients are self-selected or selected by their physicians into various treatments for reasons (disease severity, contraindications, socioeconomic status, overall prognosis, comorbidities, anticipation of adverse events, quality of life issues, coverage design, and provider preference) that are often associated with the outcomes. As the first step, researchers may have to describe and recognize the diversity in the existing treatment regimens or testing modalities in usual care. Then, a thorough understanding of how treatment selection is made in the real world is necessary for accurate definition and operationalization of "usual or standard care." Note that standards of care may vary across geographic regions and treatment settings, or may change over time. It is important to recognize that a "waste basket definition" of "usual or standard care" (any users of any existing treatments) should be avoided for the reasons mentioned above. Lastly, it is important to recognize that comparisons may be impossible when suspected or observed differences between the exposure and comparison groups are associated with the outcome of interest and cannot be adequately adjusted for and controlled through study design or analytic

approaches (i.e., in situations with intractable confounding).

Historical Comparison

A historical comparison group may seem to be a natural choice when there is a dramatic shift from one treatment to another (e.g., rapid diffusion of a new treatment in practice, or sudden change in treatment utilization due to evidence or practice changes). It may also be the only choice when there is such strong selection for the new treatment that it is uncontrollable even with rigorous methods and randomization, is unethical, or is not realistic for other reasons. However, in any situation, the use of a historical control needs to be justified after considering associated methodological issues.

Historical comparison groups will still be vulnerable to confounding by indication or severity when information on indication or severity is unmeasured. To overcome this limitation, an instrumental variable (IV) analysis using calendar time as an instrument has been applied.[8-11] Even in analyses using calendar time as an IV, confounding by indication may still arise if time is associated with severity and outcomes of interest. When historical comparison groups are used, any changes in the severity or operational definitions of the target condition as well as changes in outcome rates or outcome definitions over time could introduce bias into the analyses and must be adequately controlled. If these time-varying factors are not controllable, the use of a historical comparison group cannot be justified.

Comparison Groups From Different Data Sources

Situations may arise when the desired comparison groups are not available within the same data sources as the exposure groups. Multiple data sources can be linked to enhance the validity of observational comparative effectiveness and safety studies.[12-14] Registries have been linked to other data sources (e.g., Medicare data, HMO administrative data) to identify long-term clinical outcomes.[12-13] Although device or drug registries may provide detailed data on the use of drugs, biologics, and devices and on the severity of underlying disease and related comorbidities, registries are often limited to one product or a class of product, and therefore may not contain information on the comparison group of interest. In this situation, other existing disease, drug, or device registries have been considered to identify comparison groups.[13-14] Suppose, for example, that researchers linked a registry for a device and a separate clinical registry for the target condition to Medicare data to identify the exposure and comparison group within Medicare-linked patients. In this study, both exposure and comparison groups are obtained from the same source population (Medicare); however, sampling of each group may be different, as each registry may have collected data through a different mechanism.

At least two potential issues need to be considered when using comparison groups from different databases: (1) residual confounding and (2) generalizability (a concept related to target populations). Residual confounding could arise in comparisons across different data sources for two reasons. First, residual confounding might occur due to incomparability of information in exposure and comparison groups. It is common that information about the patient, exposure or comparison treatment, and/or outcome is collected differently across different databases, and therefore is not comparable between the exposed group and comparison group. This noncomparability of available information for confounder adjustment may lead to increased residual confounding when common variables available across the databases are limited. Second, increased residual confounding is also possible because exposed patients and comparison patients may be different in observed and unobserved domains because they are sampled differently or because they may come from a different source population.[15] In the previous example of a study using two registries linked to Medicare, it is possible that two groups are different with respect to demographic characteristics and/or geographic regions even though they are all Medicare patients. Because many factors associated with socioeconomic status that might be associated with treatment choice and outcomes are unmeasured, comparisons across different databases could cause increased residual confounding. The problem may be minimized by adequate consideration of hospital clusters and with attempts to control for surrogates for socioeconomic status.

A separate issue of generalizability could arise as estimation of a causal effect in observational studies or trials necessitates a target population[16,17] and many methods of adjusting for confounding such as standardization and inverse-probability-of-treatment-weighting are based on the idea of estimating average treatment effect in a target population.[18-19] Describing a finite population that the effect estimates would be computed for and apply to may be challenging when exposure groups and comparison groups come from different databases. In the previous example study of device and clinical registries linked to Medicare, the finite target population could be defined as Medicare patients. However, when each registry is not a random sample of Medicare patients but selects a very different sample, the generalizability of the findings from the study (assuming that residual confounding is taken care of) could be complex to understand. When using comparison groups from multiple databases, researchers need to clearly describe the methods and consider and discuss the issues outlined here to increase the validity and interpretability of their findings.

Operationalizing the Comparison Group in CER

A number of important considerations regarding the definition, measurement, and operationalization of exposure are discussed in chapter 4, and apply equally to the operationalization of comparator group(s). Below, we discuss issues that specifically affect the operationalization of the comparator(s).

Indication

As discussed, the overriding consideration that should guide comparator choice is the generation of evidence that directly informs decisions on treatments, testing, or health care delivery systems as defined in the study question. Thus, another treatment used for the same indication as the exposure treatment will typically be used as the comparison group for assessing comparative effectiveness. When a treatment and a comparison treatment have a single and specific indication, such as insulin and glitazones for diabetes, and are not commonly used off-label for other conditions, the indication may simply be inferred by the initiation of the treatment. However, because many treatments, particularly drugs, are approved for and/or clinically used to treat multiple indications, the appropriate indication will often have to be ensured by defining the indication and restricting the study population. Defining the indication typically involves requirement for the presence of certain diagnoses, the absence of diagnoses for alternative indications, or a combination of both,[20] but also depends on how the comparative effectiveness question was formulated, that is, what the target population is and whether the population is defined by indications and contraindications. It is important to recognize that restriction of the study population to patients with the same indication does not necessarily remove confounding by severity.[21]

For clinical effectiveness or safety questions, nonusers or users of other treatments (active comparators) with different indications may be considered as comparison groups. For nonuser comparisons, restriction of nonusers to those with similar indications is advisable. However, such restriction is unlikely to fully address healthy user bias, and randomization may be necessary to study such clinical effectiveness questions.[22] Active comparators, as explained in the previous section, are generally more appropriate, particularly for safety questions, and their use may reduce or eliminate healthy user bias.

Initiation

There are well-recognized advantages in studying new initiators of treatments, which is why the new user design is considered the gold standard in pharmacoepidemiology.[23] Specifically, a new user design prevents under-ascertainment of early events and avoids problems arising from confounders that may be affected by treatment in prevalent users.[23] It also prevents bias arising from prevalent users being long-term adherers who may also follow other healthy behaviors.[4, 24] See chapter 2 for a complete discussion of the new user design.

Inclusion of prevalent users may be justified, however, when outcomes of interest are extremely rare or occur after long periods of use, so that a new user design may not be feasible. The benefits and potential bias arising from the inclusion

of prevalent users should be carefully weighed, and the evidence generated by the design may be considered hypothesis generating rather than hypothesis testing. Comparisons between incident and prevalent users should be avoided. As for the exposure of interest, introduction of immortal time through incorrect classification of person-time has to be avoided for both the exposure and comparison group.[6]

Exposure Time Window

As discussed in chapter 4, each exposure group requires the definition of an exposure-time window that corresponds to the period where therapeutic benefit and/or risk would plausibly occur, and that could substantially differ from the actual exposure to the treatment.[25] Importantly, this exposure window can differ between the exposure of interest and the comparator(s), and the determination of the appropriate time window should be made individually for each group based on the pharmacologic or therapeutic profile of the intervention. Time-to-event analyses including Cox proportional hazard regression may be appropriate when comparing two treatments with expected differences in the timing of beneficial or safety outcomes.

In situations where there is uncertainty regarding the appropriate duration of the exposure window(s), sensitivity analyses should be performed to assess whether results are sensitive to different specifications of the exposure window(s). In addition, performing both an as-treated analysis (where patients are censored at the end of the exposure-time window) as well as an intention-to-treat (ITT, i.e., first-exposure-carried-forward analysis) may help understand the impact of nonadherence, misclassification, and censoring on the observed results. However, it is important to recognize that the utility of ITT analyses are generally limited when assessing long-term effects. Conversely, as-treated analyses could cause bias due to informative censoring (when stopping is associated with the outcome of interest), so methods to model and address informative censoring should be considered.[26] Comparisons between implantable devices and drug treatments present a special case of ITT analysis, as the "as treated" and ITT specifications will result in very similar exposure durations for devices (because of the inability to discontinue an implantable device other than in cases of device failure/removal), but may result in dramatically different exposure durations for drug treatments with high discontinuation rates; this must be taken into account when determining the followup periods that should be included in study analyses for both comparators.

Nonadherence

Nonadherence to prescribed medications is common and a recognized problem for the health care system. Nonadherence may be different between treatment and comparator(s) due to differences in complexity of dosing regimens, side effect profiles, and patient preferences. Because CER aims to compare benefits and harms of different interventions in real-world conditions, treatment effects should be compared at adherence levels observed in clinical practice rather than adjusting for the difference in adherence. When adherence to a comparator is lower than adherence to the exposure treatment of interest and both treatments have similar benefits when used as prescribed, the benefit of the exposure treatment will be superior due to better adherence. Since the aim of the study is estimation of drug effects in real world situation and patients, the results are valid. However, it is important to report adherence measures for each of the treatments as part of the study results so that findings can be interpreted under appropriate consideration of the observed adherence patterns. Requiring run-in periods to assure that adherence is satisfactory and more equal across groups[27] may be problematic because such practice could introduce immortal time bias (if the run-in period is included in the analysis) or be unable to estimate effects in the early phase of treatment (if the run-in period is excluded from analysis).

Dose/Intensity of Drug Comparison

After the study population has been defined and exposure and comparison groups have been chosen, it is important to appreciate the effects of dose on outcomes. When there is a dose effect on the outcome of interest, the dose of the exposure and comparison drug(s) will drive the direction and the magnitude of effects. A lower-dose comparison drug may make the study drug look more effective,

while a higher-dose comparison drug may make the study drug look safer. Therefore, researchers first should assess and report the dose in each group. When appropriate and possible, comparisons should be made for exposure and comparison group at various clinically equivalent dose levels. It is important to recognize that comparisons between different dose levels may potentially result in confounding by severity, as higher doses are likely to be given to patients with more severe disease.

Considerations for Comparisons Across Different Treatment Modalities

Many principles in the previous sections are discussed primarily in the context of medications. In this section we focus specifically on the important methodological issues for comparisons across different treatment modalities.

Confounding by Indication or Severity

For some conditions, drugs may be used for patients with a milder disease, and surgery may be reserved for those with more severe disease. In many circumstances, a step-wise approach to treat a condition may be recommended or practiced (e.g., consider a surgery if a drug treatment failed). For other diseases like cancer, early-stage disease may be treated with surgical procedures, whereas more advanced disease may be treated with chemotherapy and/or radiation or combinations of multiple modalities. Although not different from within-drug or within-procedure/surgery comparisons, understanding the recommendations from guidelines and standards of practice is necessary to assess the direction and magnitude of potential confounding by indication or severity when comparing across different treatment modalities.

Selection of Healthier Patients into More Invasive Treatments

While invasiveness of surgeries and procedures varies, they typically pose short-term risks in exchange for long-term benefits. Therefore, patients who are not in good general condition due to severe target disease or comorbidities are less likely to be considered for invasive procedures. This potential bias due to selection of healthier patients into more invasive treatment is more problematic in comparisons across different treatment modalities, especially when indications and severity are not adequately accounted for in the selection of exposure and comparison groups. Being selected for surgeries or procedures may be a surrogate for better general conditions, including having less severe disease and comorbid conditions as well as better functional and psychological well-being. Furthermore, surgery/procedures are more expensive and typically offered through specialists' care. Therefore, selection of wealthier and more health-seeking patients into surgery/procedures may be expected.

The direction of bias may be unpredictable when both confounding by indication/severity and healthy user bias come into play. In general, controlling for healthy user bias is challenging and may only be achieved in observational studies when information on health behaviors or their surrogates are available in all or a subset of patients, or a good instrument exists to allow a valid instrumental variable analysis. Sensitivity analyses assessing the impact of healthy user bias is necessary and more research is needed to understand factors associated with the selection of patients into surgery/procedures to understand the magnitude of potential healthy user bias in the device-drug comparison settings.

Time from Disease Onset to a Treatment

If not appropriately accounted for, lag times between date of initial diagnosis and date of treatment may create bias in studies assessing comparative or clinical effectiveness. For example, when assessing comparative survival after heart transplantation, there is a waiting time between referral to surgery and receipt of transplantation.[28] Currently, most patients are treated with (or bridged by) left ventricular assist devices (LVAD). Comparing the survival after LVAD to that after transplantation will be biased (i.e., immortal time bias) if researchers fail to take the sequence of these treatments into account and adequately allocate person-time on the first treatment (LVAD).

Another pertinent example of immortal person-time bias in clinical effectiveness research is the comparison of survival for responders and nonresponders to chemotherapy.[29] As responders to chemotherapy have to survive through the period of responding to chemotherapy to be identified as responders, this comparison will suffer from

"time-to-response" or immortal person-time bias if not adequately controlled.[29] This problem has recently been described by Suissa using pharmacoepidemiological examples.The same problem arises with even greater magnitude when a medical treatment is compared to a surgical treatment and patients are treated with the medical treatment prior to being referred to the surgery if surgery is considered for more advanced disease (or vice versa). Careful attention to the time from initial diagnosis and general sequence of different treatment modalities is needed to prevent immortal person-time bias.

Different Magnitude of Misclassification in Drug Exposure Versus Procedure Comparison

Assessment of drug exposure in existing data sources always requires assumptions, as longitudinal records that measure patients' actual intake of medications are not available in large databases. Pharmacy records in many administrative databases for government or commercial insurance agencies are considered the "gold standard" in pharmacoepidemiology as they capture longitudinal pharmacy dispensing in a large number of subjects. However, pharmacy dispensing does not provide information on the actual intake of medications by patients, and most drug exposure is chronic rather than acute. Therefore, defining drug exposure using dispensing data requires certain assumptions and some degree of exposure misclassification is always expected. On the other hand, assessment of exposure to surgery or procedure (especially major procedures that are well reimbursed or clinically important) is more straightforward, and their identification is likely to be less affected by misclassification as these one-time or acute major clinical events are usually accurately recorded in administrative databases or registries. When comparing drug exposures with surgeries or procedures, researchers need to recognize that misclassification is likely not comparable in both groups, and they need to assess how this potential misclassification affects their results.

Provider Effects in Devices or Surgeries

Characteristics of the operating physician and institution where the device implantation or surgery was carried out are important factors to consider when evaluating the comparative effectiveness of medical devices or surgeries. Certain physician and institutional characteristics such as experience and specialty are known to affect outcomes, particularly during the periprocedural period. A direct relationship between level of physician experience and better patient outcomes has been documented for technically complex procedures and implantations like angioplasty, stenting, and various surgeries.[30-33] A relationship between larger hospital volume and favorable patient outcomes for a variety of procedures is also well documented.[31-32, 34-38] While these factors are more likely to behave as confounders than as effect measure modifiers, stratification must first be carried out to inform decisions on how to handle these factors.Therefore, it is necessary to be able to identify physicians and institutes for a device implantation or surgery and characteristics such as volume of procedures that are known to affect outcomes. In addition, exploring physician effects in the study population to account for provider effects is necessary to conduct valid comparisons including devices or surgeries.

Adherence to Drugs and Device Failure or Removal

Patients who are on medications could have various degrees of adherence, from completely stopping, skipping doses, to taking medications as prescribed. Measuring adherence is not impossible but requires assumptions in most data sources. On the other hand, implantable devices or surgical procedures do not generally have adherence issues unless there is a device failure or a complication that requires device removal. For most implantable devices, removal is a major procedure and therefore likely to be captured accurately. However, a unique problem could arise for devices with a function to be turned off (without being removed). How to take adherence and device failure or removal into account depends on the goal of each study and how the researchers define effectiveness. If the goal is to assess effectiveness in real-world patients and practice where nonadherence is common and some degree of device failure or removal is expected, simply comparing two different modalities without adjusting for adherence or device failure should be appropriate. It is recommended that both adherence and device failure rates are assessed and reported. However, if

the goal is to compare the conditional effectiveness assuming perfect adherence or no device failure, the question should be clearly stated and the appropriate design and/or method for adjustment needs to be employed.

Conclusion

Understanding the impact of comparator choice on study design is important when conducting observational CER. While this choice affects the potential for and magnitude of confounding and other types of bias, the selection of a comparator group should be primarily driven by a comparative effectiveness question that has been prioritized by the informational need of the stakeholder community. The overriding consideration that should guide comparator choice is the generation of evidence that directly informs decisions on treatments, testing, or health care delivery systems as defined by the study question. Researchers engaged in observational CER need to keep in mind that there may be questions (comparisons) not validly answered due to intractable bias in observational CER.

Checklist: Guidance and key considerations for comparator selection for an observational CER protocol

Guidance	Key Considerations	Check
Choose concurrent, active comparators from the same source population (or justify use of no-treatment comparisons/historical comparators/ different data sources).	- Comparator choice should be primarily driven by a comparative effectiveness question prioritized by informational needs of the stakeholder community and secondarily as a strategy to minimize bias.	☐
Discuss potential bias associated with comparator choice and methods to minimize such bias, when possible.	- Be sure to also describe how study design/analytic methods will be used to minimize bias.	☐
Define time zero for all comparator groups in describing planned analyses.	- Choice of time zero, particularly in no-treatment or usual care, should be carefully considered in light of potential immortal person-time bias and prevalent user bias. - Employ a new user design as a default, if possible.	☐

References

1. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. Mar 2006;98(3):253-9.
2. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf. May 2006;15(5):291-303.
3. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. N Engl J Med. May 21, 1998;338(21):1516-20.
4. Setoguchi S, Glynn RJ, Avorn J, et al. Statins and the risk of lung, breast, and colorectal cancer in the elderly. Circulation. Jan 2, 2007;115(1):27-33.
5. Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology. Nov 2001;12(6):682-9.
6. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. Feb 15, 2008;167(4):492-9.
7. Pencina MJ, Larson MG, D'Agostino RB. Choice of time scale and its effect on significance of predictors in longitudinal studies. Statistics in Medicine. 2007;26(6):1343-59.
8. Cain LE, Cole SR, Greenland S, et al. Effect of highly active antiretroviral therapy on incident AIDS using calendar period as an instrumental variable. Am J Epidemiol. May 1, 2009;169(9):1124-32.

9. Johnston KM, Gustafson P, Levy AR, et al. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. Statistics in Medicine. 2008;27(9):1539-56.
10. Rascati KL, Johnsrud MT, Crismon ML, et al. Olanzapine versus risperidone in the treatment of schizophrenia: A comparison of costs among Texas Medicaid recipients. PharmacoEconomics. 2003;21(10):683-97.
11. Shetty KD, Vogt WB, Bhattacharya J. Hormone replacement therapy and cardiovascular health in the United States. Med Care. May 2009;47(5): 600-6.
12. Hernandez AF, Fonarow GC, Hammill BG, et al. Clinical effectiveness of implantable cardioverter-defibrillators among Medicare beneficiaries with heart failure. Circulation: Heart Failure. January 1, 2010;3(1):7-13.
13. Setoguchi S. AHRQ Effective Health Care Program Ongoing Study: Real World Effectiveness of Implantable Cardioverter Defibrillators (ICDs) in Medicare Patients. 2010. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=431. Accessed May 24, 2011.
14. Askling J, van Vollenhoven RF, Granath F, et al. Cancer risk in patients with rheumatoid arthritis treated with anti–tumor necrosis factor α therapies: Does the risk change with the time since start of treatment? Arthritis & Rheumatism. 2009;60(11):3180-9.
15. Hammill B, Curtis LH, Setoguchi S. Performance of propensity score methods when comparison groups originate from different data sources. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 2:81-9.
16. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol. April 1, 2002;31(2):422-9.
17. Shahar E. Estimating causal parameters without target populations. J Eval Clin Pract. 2007;13(5):814-6.
18. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550-60.
19. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003;14(6):680-6.
20. Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Med Care. Oct 2007;45(10 Supl 2):S131-42.
21. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. Am J Epidemiol. 1999 Jun 1;149(11):981-3.
22. Shrank WH, Patrick AR, Brookhart MA. Healthy user and related biases in observational studies of preventive interventions: a primer for physicians. J Gen Intern Med. 2011 May;26(5):546-50. Epub 2011 Jan 4.
23. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. Nov 1, 2003;158(9):915-20.
24. Setoguchi SA, Schneeweiss S. Statins and the Risk of Colorectal Cancer. N Engl J Med. 2005;353(9):952-4.
25. van Staa TP, Abenhaim L, Leufkens H. A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. J Clin Epidemiol. Feb 1994;47(2):183-9.
26. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. Sep 2004;15(5):615-25.
27. Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: Approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: The International Society for Pharmacoeconomics and Value in Health (Wiley-Blackwell). 2009;12(8):1053-61.
28. Mantel N, Byar DP. Evaluation of response-time data involving transient states: An illustration using heart-transplant data. Journal of American Statistical Association. 1974;69:81-6.
29. Anderson J, Cain K, Gelber R. Analysis of survival by tumor response. J Clin Oncol. November 1, 1983;1(11):710-9.
30. Jollis JG, Peterson ED, Nelson CL, et al. Relationship between physician and hospital coronary angioplasty volume and outcome in elderly patients. Circulation. Jun 3, 1997;95(11):2485-91.

31. McGrath PD, Wennberg DE, Dickens JD, Jr., et al. Relation between operator and hospital volume and outcomes following percutaneous coronary interventions in the era of the coronary stent. JAMA. Dec 27, 2000;284(24):3139-44.

32. Hannan EL, Racz M, Ryan TJ, et al. Coronary angioplasty volume-outcome relationships for hospitals and cardiologists. JAMA. Mar 19, 1997;277(11):892-8.

33. Birkmeyer JD, Stukel TA, Siewers AE, et al. Surgeon volume and operative mortality in the United States. N Engl J Med. Nov 27, 2003;349(22):2117-27.

34. Luft HS, Bunker JP, Enthoven AC. Should operations be regionalized? The empirical relation between surgical volume and mortality. N Engl J Med. Dec 20, 1979;301(25):1364-9.

35. Showstack JA, Rosenfeld KE, Garnick DW, et al. Association of volume with outcome of coronary artery bypass graft surgery. Scheduled vs nonscheduled operations. JAMA. Feb 13, 1987;257(6):785-9.

36. Cebul RD, Snow RJ, Pine R, et al. Indications, outcomes, and provider volumes for carotid endarterectomy. JAMA. Apr 22-29, 1998;279(16):1282-7.

37. Urbach DR, Baxter NN. Does it matter what a hospital is "high volume" for? Specificity of hospital volume-outcome associations for surgical procedures: analysis of administrative data. Qual Saf Health Care. Oct 2004;13(5):379-83.

38. Birkmeyer JD, Siewers AE, Finlayson EV, et al. Hospital volume and surgical mortality in the United States. N Engl J Med. Apr 11, 2002;346(15):1128-37.