Chapter 11. Sensitivity Analysis

Abstract

This chapter provides an overview of study design and analytic assumptions made in observational comparative effectiveness research (CER), discusses assumptions that can be varied in a sensitivity analysis, and describes ways to implement a sensitivity analysis. All statistical models (and study results) are based on assumptions, and the validity of the inferences that can be drawn will often depend on the extent to which these assumptions are met. The recognized assumptions on which a study or model rests can be modified in order to assess the sensitivity, or consistency in terms of direction and magnitude, of an observed result to particular assumptions. In observational research, including much of comparative effectiveness research, the assumption that there are no unmeasured confounders is routinely made, and violation of this assumption may have the potential to invalidate an observed result. The analyst can also verify that study results are not particularly affected by reasonable variations in the definitions of the outcome/exposure. Even studies that are not sensitive to unmeasured confounding (such as randomized trials) may be sensitive to the proper specification of the statistical model. Analyses are available that can be used to estimate a study result in the presence of an hypothesized unmeasured confounder, which then can be compared to the original analysis to provide quantitative assessment of the robustness (how much does the estimate change if we posit the existence of a confounder) of the original analysis to violations of the assumption of no unmeasured confounders. Finally, an analyst can examine whether specific sub-populations should be addressed in the results since the primary results may not generalize to all sub-populations if the biologic response or exposure may differ in these subgroups. The chapter concludes with a checklist of key considerations for including sensitivity analyses in a CER protocol or proposal.

Introduction

Observational studies and statistical models rely on assumptions, which can range from how a variable is defined or summarized, to how a statistical model is chosen and parameterized. Often these assumptions are reasonable, and, even when violated, may result in unchanged effect estimates. When the results of analyses are consistent or unchanged by testing variations on underlying assumptions, they are said to be "robust." However, violations in assumptions that result in meaningful effect estimate changes provide insight into the validity of the inferences that might be drawn from a study. A study's underlying assumptions can be altered along a number of dimensions, including study definitions (modifying exposure/outcome/confounder definitions), study design (changing or augmenting the data source or population under study), and modeling (modifying a variable's functional form or testing normality assumptions), to evaluate robustness of results.

This chapter considers the forms of sensitivity analysis that can be included in the analysis of an observational comparative effectiveness study, provides examples, and offers recommendations about the use of sensitivity analyses.

Unmeasured Confounding and Study Definition Assumptions

Unmeasured Confounding

The underlying assumption of all epidemiological studies is that there is no unmeasured confounding. However, some potential confounding variables may not be measured or available for analysis: the unmeasured confounding variable could either be a known confounder that is not present in the type of data being used (e.g., obesity is commonly not available in prescription claims databases) or an unknown confounder where the confounding relation is unsuspected. Quantifying the effect that an unmeasured confounding variable would have on study results provides an assessment of the sensitivity of the result to violations of the assumption of no unmeasured confounding. The robustness of an association to the presence of a confounder[1,2] can alter inferences that might be drawn from a study, which then might change how the study results are used to influence translation into clinical or policy decision-making. Methods for assessing the potential impact of unmeasured confounding on study results, as well as quasi-experimental methods to account for unmeasured confounding, are discussed later in the chapter.

Comparison Groups

An important choice in study design is the selection of suitable treatment and comparison groups. This step can serve to address many potential limitations of a study, such as how new user cohorts eliminate the survivor bias that may be present if current (prevalent) users are studied (which would reflect only people who could tolerate the treatment and, most likely, for whom treatment appeared to be effective).[3] However, this new user approach can limit the questions that can be asked in a study, as excluding prevalent users might omit long term users (which could overlook risks that arise over long periods of use). For example, when Rietbrock *et al.* considered the comparative effectiveness of warfarin and aspirin in atrial fibrillation[4] in the General Practice Research Database, they looked at current use and past use instead of new use. This is a sensible strategy in a general practice setting as these medications may be started long before the patient is diagnosed with atrial fibrillation. Yet, as these medications may be used for decades, long term users are of great interest. In this study, the authors used past use to address indication, by comparing current users to past users (an important step in a prevalent users study).

One approach is to include several different comparison groups and use the observed differences in potential biases with the different comparison groups as a way to assess the robustness of the results. For example, when studying the association between thiazide diuretics and diabetes, one could create reference groups including “non-users”, “recent past-users”, “distant past-users” and “users of other anti-hypertensive medications”. One would presume that the risk of incident diabetes among the “distant past-users” should resemble that of the “non-users”; if not there is a possibility that confounding by indication is the reason for the difference in risk.

Exposure Definitions

Establishing a time window that appropriately captures exposure during etiologically relevant time periods can represent a challenge in study design when decisions need to be made in the presence of uncertainty.[5] Uncertainty about the most appropriate way to define drug exposure can lead to questions about what would have happened if the exposure had been defined a different way. A substantially different exposure-outcome association observed under different definitions of exposure (such as different time windows or dose [e.g., either daily or cumulative]) might provide insight into the biological mechanisms underlying the association or provide clues

about potential confounding or unaddressed bias. As such, varying the exposure definition and re-analyzing under different definitions serves as a form of sensitivity analysis.

Outcome Definitions

The association between exposure and outcome can also be assessed under different definitions of the outcome. Often a clinically relevant outcome in a data source can be ascertained in several ways (e.g., a single diagnosis code, multiple diagnosis codes, a combination of diagnosis and procedure codes, etc.). The analysis can be repeated using these different definitions of the outcome, which may shed light on the how well the original outcome definition truly reflects the condition of interest.

Beyond varying a single outcome definition, it is also possible to evaluate the association between the exposure and clinically different outcomes. If the association between the exposure and one clinical outcome is known from a study with strong validity (such as from a clinical trial) and can be reproduced in the study, the observed association between the exposure of interest and an outcome about which external data are not available becomes more credible. Since some outcomes might not be expected to occur immediately after exposure (e.g., cancer), the study could employ different lag (induction) periods between exposure and the first outcomes to be analyzed in order to assess the sensitivity of the result to the definition. This result can lead either to insight into potential unaddressed bias or confounding, or it could be used as a basis for discussion about etiology (e.g., does the outcome have a long onset period).

Covariate Definitions

Covariate definitions can also be modified to assess how well they address confounding in the analysis. Although a minimum set of covariates may be used to address confounding, there may be an advantage to using a staged approach where groups of covariates are introduced, leading to progressively greater adjustment. If done transparently, this approach may provide insight into which covariates have relatively greater influences on effect estimates, permitting comparison with known or expected associations or permitting the identification of possible intermediate variables.

Finally, some covariates are known to be misclassified under some approaches. A classic example is an “intention to treat” analysis which assumes that each participant continues to be exposed once they have received an initial treatment. Originally used in the analysis of randomized trials, this approach has been used in observational studies as well.[6] It can be worthwhile to do a sensitivity analysis on studies that use an intention to treat approach to see how different an “as treated” analysis would be even if intention to treat is the main estimate of interest, mostly in cases where there is differential adherence in the data source between two therapeutic approaches.[7]

Summary Variables

Study results can also be affected by the summarization of variables. For example, time can be summarized, and differences in the time window during which exposure is determined can lead to changes in study effect estimates. For example, the risk of venous thromboembolism rises with duration of use for oral contraceptives;[8] an exposure definition that did not consider the cumulative exposure to the medication might underestimate the difference in risk between two

different formulations of oral contraceptive. Alternately, effect estimates may vary with changes in the outcome definition. For example, an outcome definition of all cardiovascular events including angina could lead to a different effect estimate than an outcome definition including only myocardial infarction. Sensitivity analyses of the outcome definition can allow for a richer understanding of the data, even for models based on data from a randomized controlled trial.

Selection Bias

The assessment of selection bias through sensitivity analysis involves assumptions regarding inclusion or participation by potential subjects, and results can be highly sensitive to assumptions. For example, the over-sampling of cases exposed to one of the drugs under study (or, similarly, an under-sampling) can lead to substantial changes in effect measures over ranges that might plausibly be evaluated. Even with external validation data, which may work for unmeasured confounders,[9] it is difficult to try and account for more than a trivial amount of selection bias. Generally, if there is strong evidence of selection bias in a particular data set it is best to seek out alternative data sources.

One limited exception may be when the magnitude of bias is known to be small.[10] This may be true for non-random loss to follow-up in a patient cohort. Since the baseline characteristics of the cohort are known, it is possible to make reasonable assumptions about how influential this bias can be. But, in the absence of such information, it is generally better to focus on identifying and eliminating selection bias at the data acquisition or study design stage.

Data Source, Sub-Populations, and Analytic Methods

The first section of this chapter covered traditional sensitivity analysis to test basic assumptions such as variable definitions and to consider the impact of an unmeasured confounder. These issues should be considered in every observational study of comparative effectiveness research. However, there are some additional sensitivity analyses that should be considered, depending on the nature of the epidemiological question and the data available. Not every analysis can (or should) consider these factors, but they can be as important as the more traditional sensitivity analysis approaches.

Data Source

For many comparative effectiveness studies, the data used for the analysis were not specifically collected for the purpose of the research question. Instead, the data may have been obtained as part of routine care or for purposes of medical billing. In such cases, it may be possible to acquire multiple data sources for a single analysis (and use the additional data sources as a sensitivity analysis). Where this is not feasible, it may be possible to consider differences between study results and results obtained from other papers that use different data sources. While all data sources have inherent limitations in terms of the data that are captured by the database, these limitations can be accentuated when the data were not prospectively collected for the specific research purpose.[11] For example, secondary use of data increases the chances that a known but unmeasured confounder may explain part or all of an observed association. A straightforward example of the differences in data capture can be seen by comparing data from Medicare (i.e., US medical claims data) and the General Practice Research Database (i.e., British electronic medical records collected as part of routine care).[11] Historically, Medicare data have lacked the results of routine laboratory testing and measurement (quantities like height, weight,

blood pressure, and glucose measures), but include detailed reporting on hospitalizations (which are billed and thus well recorded in a claims database). In a similar sense, historically, the General Practice Research Database has had weaker reporting on hospitalizations (since this information is captured only as reports given back to the General Practice, that usually are less detailed), but better recording than Medicare data for routine measurements (such as blood pressure) that are done as part of a standard medical visit.

Issues with measurement error also can emerge because of the process by which data are collected. For example, "myocardial infarction" coded for the purposes of billing may vary slightly or substantially from a clinically verified outcome of myocardial infarction. As such, there will be an inevitable introduction of misclassification into the associations. Replicating associations in different data sources can provide an idea of how changes to the operational definition of an outcome can alter the estimates (e.g., comparing a report to a general practitioner (GP) with a hospital ICD-9 code). Replication of a study using different data sources is more important for less objectively clear outcomes (such as depression) than it is for more objectively clear outcomes (such as all-cause mortality).

An analysis conducted in a single data source may be vulnerable to bias due to systematic measurement error or the omission of a key confounding variable. Associations that can be replicated in a variety of data sources, each of which may have used different definitions for recording information and which have different covariates available, provide reassurance that the results are not simply due to the unavailability of an important confounding variable in a specific data set. Furthermore, when estimating the possible effect of an unmeasured confounder on study results, data sets that measure the confounder may provide good estimates of the confounder's association with exposure and outcome (and provide context for results in data sources without the same confounder information).

An alternative to looking at completely separate datasets is to consider supplementing the available data with external data sources. An example of a study that took the approach of supplementing data was conducted by Huybrechts *et al.*[12]. They looked at the comparative safety of typical and atypical antipsychotics among nursing home residents. The main analysis used prescription claims (Medicare and Medicaid data) and found, using high dimensional propensity score adjustment, that conventional anti-psychotics were associated with an increase in 180 day mortality risk (a risk difference of 7.0 per 100 persons [95% CI: 5.8, 8.2]). The authors then included data from MDS (Minimum Data Set) and OSCAR (Online Survey, Certification and Reporting) which contains clinical covariates and nursing home characteristics.[12] The result of including these variables was an essentially identical estimate of 7.1 per 100 persons (95% CI: 5.9, 8.2).[12] This showed that these differences were robust to the addition of these additional covariates. It did not rule out other potential biases, but it did demonstrate that simply adding MDS and OSCAR data would not change statistical inference.

While replicating results across data sources provides numerous benefits in terms of understanding the robustness of the association and reducing the likelihood of a chance finding, it is often a luxury that is not available for a research question, and inferences may need to be drawn from the data source at hand.

Key Sub-Populations

Therapies are often tested on an ideal population (e.g., uncomplicated patients thought to be likely to adhere to medication) in clinical trials. Once the benefit is clearly established in trials, the therapy is approved for use and becomes available to all patients. However, there are several cases where it is possible that the effectiveness of specific therapies can be subject to effect measure modification. While a key sub-population may be independently specified as a population of interest, showing that results are homogeneous across important sub-populations can build confidence in applying the results uniformly to all sub-populations. Alternatively, it may highlight the presence of effect measure modification and the need to comment on population heterogeneity in the interpretation of results. As part of the analysis plan, it is important to state whether measures of effect will be estimated within these or other sub-populations present in the research sample in order to assess possible effect measure modification:

Pediatric populations. Children may respond differently to therapy than adults, and dosing may be more complicated. Looking at children as a separate and important sub-group may make sense if a therapy is likely to be used in children.

Genetic variability. The issue of genetic variability is often handled only by looking at different ethnic or racial groups (who are presumed to have different allele frequencies). Some medications may be less effective in some populations due to the different polymorphisms that are present in these persons, though indicators of race and ethnicity are only surrogates for genetic variation.

Complex patients. These are patients who suffer from multiple disease states at once. These disease states (or the treatment[s] for these disease states) may interfere with each other, resulting in a different optimal treatment strategy in these patients. A classic example is the treatment of cardiovascular disease in HIV infected patients. The drug therapy used to treat the HIV infection may interfere with medication intended to treat cardiovascular disease. Treatment of these complex patients is of great concern to clinicians, and these patients should be considered separately where sample size considerations allow for this.

Older adults. Older adults are another population that may have more drug side effects and worse outcomes from surgeries and devices. Furthermore, older adults are inherently more likely to be subject to polypharmacy and thus have a much higher risk of drug-drug interactions.

Most studies lack the power to look at all of these different populations, nor are they likely to be all present in a single data source. However, when it is feasible to do so, it can be useful to explore these sub-populations to determine if the overall associations persist or if the best choice of therapy is population dependent. These can be important clues in determining how stable associations are likely to be across key sub-populations. In particular, the researcher should identify segments of the population to which there are concerns about generalizing results. For example, randomized trials of heart failure often exclude large portions of the patient population due to complexity of the underlying disease state.[13] It is critical to try to include inferences to these complex sub-populations when doing comparative effectiveness research with heart failure as the study outcome, as that is precisely where the evidence gap is the greatest.

Cohort Definition and Statistical Approaches

If it is possible to do so, it can also be extremely useful to consider the use of more than one cohort definition or statistical approach to ensure that the effect estimate is robust to the assumptions behind these approaches. There are several options to consider as alternative analysis approaches.

Samy Suissa illustrated how the choice of cohort definition can affect effect estimates in his paper on immortal time bias.[14] He considered five different approaches to defining a cohort, with person time incorrectly allocated (causing immortal time bias) and then repeated these analyses with person time correctly allocated (giving correct estimates). Even in this straightforward example, the corrected hazard ratios varied from 0.91 to 1.13 depending on the cohort definition. There were five cohort definitions used to analyze the use of anti-thrombotic medication and the time to death from lung cancer: time-based cohort, event-based cohort, exposure-based cohort, multiple-event-based cohort, and event-exposure-based cohort. These cohorts produce hazard ratios of 1.13, 1.02, 1.05, 0.91, and 0.95, respectively. While this may not seem like an extreme difference in results, it does illustrate the value of using varying assumptions to hone in on an understanding of the stability of the associations under study with different analytical approaches, as in this example where point estimates varied by about +/- 10% depending in how the cohort was defined.

One can also consider the method of covariate adjustment to see if it might result in changes in the effect estimates. One option to consider as an adjunct analysis is the use of a high dimensional propensity score,[15] as this approach is typically applicable to the same data upon which a conventional regression analysis is performed. The high dimensional propensity score is well suited to handling situations in which there are multiple weak confounding variables. This is a common situation in many claims database contexts, where numerous variables can be found that are associated (perhaps weakly) with drug exposure, and these same variables may be markers for (i.e., associated with) unmeasured confounders. Each variable may represent a weak marker for an unmeasured confounder, but collectively (such as through the high dimensional propensity score approach) their inclusion can reduce confounding from this source. This kind of propensity score approach is a good method for validating the results of conventional regression models.

Another option that can be used, when the data permit it, is an instrumental variable (IV) analysis to assess the extent of bias due to unmeasured confounding (see Chapter 10 for a detailed discussion of IV analysis).[16] While there have been criticisms that use of instruments such as physician or institutional preference may have assumptions that are difficult to verify and may increase the variance of the estimates[17], an instrumental variable analysis has the potential to account for unmeasured confounding factors (which is a key advantage), and traditional approaches also have unverifiable assumptions. Also, estimators resulting from the IV analysis may differ from main analysis estimators (see Identification Supplement), and investigators should ensure correct interpretation of results using this approach.

DRAFT

Examples of Sensitivity Analysis of Analytic Methods

Sensitivity analysis approaches to varying analytic methods have been used to build confidence in results. One example is a study by Schneeweiss et al.[18] of the effectiveness of aminocaproic acid compared with aprotinin for the reduction of surgical mortality during coronary-artery bypass grafting (CABG). In this study, the authors demonstrated that three separate analytic approaches (traditional regression, propensity score, and physician preference instrumental variable analyses) all showed an excess risk of death among the patients treated with aprotinin (estimates ranged from a relative risk of 1.32 [propensity score] to a relative risk of 1.64 [traditional regression analysis]). Showing that different approaches, each of which used different assumptions, all demonstrated concordant results was further evidence that this association was robust.

Sometimes a sensitivity analysis can reveal a key weakness in a particular approach to a statistical problem. Delaney et al.[19] looked at the use of case-crossover designs to estimate the association between warfarin use and bleeding in the General Practice Research Database. They compared the case-crossover results to the case-time-control design, the nested case control design, and to the results of a meta-analysis of randomized controlled trials. The case-crossover approach, where individuals serve as their own controls, showed results that differed from other analytic approaches. For example, the case crossover design with a lagged control window (a control window that is placed back one year) estimated a rate ratio of 1.3 (95% confidence interval: 1.0–1.7) compared with a rate ratios of 1.9 for the nested case control design, 1.7 for the case time control design and 2.2 for a meta-analysis of clinical trials.[18] Furthermore, the results showed a strong dependence on the length of the exposure window (ranging from a rate ratio of 1.0 to 3.6), regardless of overall time on treatment. These results provided evidence that results from a case crossover approach in this particular situation needed a cautious interpretation as different approaches were estimating incompatible magnitudes of association, were not compatible with the estimates from trials, and likely violated an assumption of the case crossover approach (transient exposure). Unlike the Schneeweiss *et al.* example,[18] for which the results were consistent across analytic approaches, divergent results require careful consideration of which approach is the most appropriate (given the assumptions made) for drawing inferences, and investigators should provide a justification for the determination in the discussion.

Sometimes the reasons for differential findings with differences in approach can be obvious (concerns over the appropriateness of the case-crossover approach, in the Delaney *et al.* example above).[18] In other cases, differences can be small and the focus can be on the overall direction of the inference (like in the Suissa example above).[14] Finally, there can be cases where two different approaches (e.g., an IV approach and a conventional analysis) yield different inferences and it can be unclear which one is correct. In such a case, it is important to highlight these differences, and to try and determine which set of assumptions makes sense in the structure of the specific problem.

Table 11.1. Study aspects that can be evaluated through sensitivity analysis

Aspect	Evaluable through Sensitivity Analysis	Further Requirements

Aspect	Evaluable through Sensitivity Analysis	Further Requirements
Confounding I: Unmeasured	Maybe	Assumptions involving prevalence, strength and direction of unmeasured confounder
Confounding II: Residual	Maybe	Knowledge/assumption of which variables are not fully measured
Selection Bias is not present	No (Maybe; Generally not testable for most forms of selection bias but some exceptions [non-random loss to follow-up] may be testable with assumptions)	Assumption or external information on source of selection bias
Missing Data	No	Assumption or external information on mechanism for missing data
Data Source	Yes	Access to additional data sources
Sub-populations	Yes	Identifier of sub-population
Statistical Method	Yes	None
Misclassification I: Covariate Definitions	Yes	None
Misclassification II: Differential misclassification	Maybe	Assumption or external information about mechanism of misclassification
Functional Form	Yes	None

Statistical Assumptions

The guidance in this section primarily focuses on studies with a continuous outcome, exposure, or confounding factor variable. Many pharmacoepidemiological studies are conducted within a claims database environment where the number of continuous variables is limited (often only age is available), and these assumptions do not apply in these settings. However, studies set in electronic medical records or in prospective cohort studies may have a wider range of continuous variables, and it is important to ensure that they are modeled correctly.

Covariate and Outcome Distributions

It is common to enter continuous parameters as linear covariates in a final model (whether that model is linear, logistic, or survival). However, there are many variables where the association with the outcome may be better represented as a transformation of the original variable.

A good example of such a variable is net personal income, a variable that is bounded at zero but for which there may be a large number of plausible values. The marginal effect of a dollar of income may not be linear across the entire range of observed incomes (an increase of $5,000 may mean more to individuals with a base income of $10,000 than those with a base income of $100,000). As a result, it can make sense to look at transformations of the data into a more meaningful scale.

DRAFT

The most common option for transforming a continuous variable is to create categories (e.g., quintiles derived from the data set or specific cut points). This approach has the advantages of simplicity and transparency, as well as being relatively non-parametric. However, unless the cut-points have clinical meaning, they can make studies difficult to compare to one another (as each study may have different cut points). Furthermore, transforming a continuous variable into a discrete form always results in loss of information that it is better to avoid, if possible. Another option is to consider transforming the variable to see if this influences the final results. The precise choice of transformation requires knowledge of the distribution of the covariate. For confounding factors, it can be helpful to test several transformations and to see the impact of the reduction in skewness (and whether a linear approximation remains appropriate).

Functional Form

The "functional form" is the assumed mathematical association between variables in a statistical model. There are numerous potential variations in functional form that can be the subject of a sensitivity analysis. Examples include the degree of polynomial expressions, splines, or additive rather than multiplicative joint effects of covariates in the prediction of both exposures and outcomes. In all of these cases, the "functional form" is the assumed mathematical association between variables, and sensitivity analyses can be employed to evaluate the effect of different assumptions. In cases where non-linearity is suspected (i.e., a non-linear relationship between a dependent and independent variable in a model), it can be useful to test the addition of a square term to the model (i.e., the pair of covariates age + age^2 as the functional form of the independent variable age). If this check does not influence the estimate of the association, then it is unlikely that there is any important degree of non-linearity. If there is an impact on the estimates for this sort of transformation, it can make sense to try a more appropriate model for the non-linear variable (such as a spline or a generalized additive model).

Transformations should be used with caution when looking at the primary exposure as they can be susceptible to overfit. Overfit occurs when you are fitting a model to random variations in the data (i.e., noise) rather than to the underlying relation; polynomial-based models are susceptible to this sort of problem. However, if one is assessing the association between a drug and an outcome, this can be a useful way to handle parameters (like age) that will not be directly used for inference but that one wishes to balance between two exposure groups. These transformations should also be considered as possibilities in the creation of a probability of treatment model (for a propensity score analysis).

Special Cases

Another modeling challenge for epidemiologic analysis and interpretation is when there is a mixture of informative null values (zeros) and a distribution. This occurs with variables like coronary artery calcium (CAC), which can have values of zero or a number of Agatston units.[20] These distributions are best modeled as two parts: one, as a dichotomous variable to determine the presence of absence of CAC; and two, using a model to determine the severity of CAC among those with CAC>0. In the specific case of CAC, the severity model is typically log transformed due to extreme skew.[Error! Bookmark not defined.] These sorts of distributions are rare, but

one should still consider the distribution and functional form of key continuous variables when they are available.

Implementations

There are a number of approaches to conducting sensitivity analyses. This section describes two widely used approaches, spreadsheet-based and code-based analyses, and is not intended to be a comprehensive guide to implementing sensitivity analyses. Other approaches to conducting sensitivity analysis exist and may be more useful for specific problems.[2]

Spreadsheet-Based

The robustness of a study result to an unmeasured confounding variable can be assessed quantitatively using a standard spreadsheet (available at http://www.drugepi.org/dope-downloads/#Sensitivity Analysis).[21] The observed result and ranges of assumptions about an unmeasured confounder (prevalence, strength of association with exposure, and strength of association with outcome) are entered into the spreadsheet and are used to provide the departure from the observed result to be expected if the unmeasured confounding variable could be accounted for using standard formulae for confounding.[22] Two approaches are available within the spreadsheet: 1) an "array" approach; and 2) a "rule out" approach. In the array approach, an array of values (representing the ranges of assumed values for the unmeasured variable) is the input for the spreadsheet. The resulting output is a three dimensional plot that illustrates, through a graphed response surface, the observed result for a constellation of assumptions (within the input ranges) about the unmeasured confounder.

In the rule out approach, the observed association and characteristics of the unmeasured confounder (prevalence and strength of association with both exposure and outcome) are entered into the spreadsheet. The resulting output is a two-dimensional graph that plots, given the observed association, the ranges of unmeasured confounder characteristics that would result in a null finding. In simpler terms, the rule-out approach quantifies, given assumptions, how strong a measured confounder would need to be to result in a finding of no association and "rules out" whether an unmeasured confounder can explain the observed association.

Statistical Software-Based

For some of the approaches discussed, the software is available online. For example, the high dimensional propensity score and related documentation is available at http://www.hdpharmacoepi.org/download/. For other approaches, like the case-crossover design[18] the technique is well known and widely available. Finally, many of the most important forms of sensitivity analysis require data management tasks (such as recoding the length of an exposure time window) that are straightforward, though time-consuming.

This section describes a few examples of handling slightly more complex functional forms of covariates (where the association is not well described by a line or by the log transformation of a line). The first example introduces a spline into a model where the analyst suspects that there might be a non-linear association with age (and there is a broad age range in the cohort, making a

linearity assumption suspect). The second example looks at how to model CAC, which is an outcome variable with a complex form.

Example of Functional Form

This SAS code is an example of a mixed model that is being used to model the trajectory of a biomarker over time (variable = years), conditional on a number of covariates. The example estimates the association between different statin medications with this biomarker. Like in many prescription claims databases, most of the covariates are dichotomous. However, there is a concern that age may not be linearly associated with outcome, so a version of the analysis is tried in which a spline is used in place of a standard age variable.

Original Analysis (SAS 9.2):

```
proc glimmix data=MY_DATA_SET;
class patientid;
model biomarker_value =age female years statinA statinB diabetes hypertension / s cl;
random intercept years/subject=patientid;
run;
```

Sensitivity Analysis:

```
proc glimmix data=MY_DATA_SET;
class patientid;
effect spl = spline(age);
model biomarker_value =spl female years statinA statinB diabetes hypertension / s cl;
random intercept years/subject=patientid;
run;
```

While the spline version of the age variable needs to be graphically interpreted, it should handle any non-linear association between age and the biomarker of interest.

Example of Two-Stage Models for Coronary Artery Calcium (CAC)

CAC is an example of a continuous variable with an extremely complex form. The examples of two-stage CAC modeling (below) use variables from the Multi-Ethnic Study of Atherosclerosis. Here, the example is testing whether different forms of non-steroidal anti-inflammatory drugs (below as asa1c, nsaid1c, cox21c) are associated with more or less calcification of the arteries. The model needs to be done in two stages as it is thought that the covariates that predict the initiation of calcification may differ from those that predict how quickly calcification progresses once the process has begun.**Error! Bookmark not defined.**

First, a model is developed for the relative risk of having a CAC score greater than zero (i.e., that there is at least some evidence of plaques looking at a CT scan of the participant's coronary arteries). The variable for CAC is cac (1=CAC present, 0=CAC not present). The repeated statement is used to invoke robust confidence intervals (as there is only one subject for each unique participant ID number, designated as the variable idno).

SAS 9.2 code example:

```
proc genmod data = b descending;
 class idno race1;
 model cac=age1c male bmi1c race1
     male  diabetes smoker ex_smoker  sbp1c dbp1c
   hdl1 ldl1 TRIG1STTN1C  asa1c nsaid1c cox21c
   / dist = poisson link = log;
 repeated subject = idno/ type =ind;
estimate 'asa1c' asa1c 1 -1/ exp;
estimate 'nsaid1c' nsaid1c 1 -1/ exp;
estimate 'cox21c' cox21c 1 -1/ exp;;
run;
```

Among those participants with CAC (as measured by an Agatston score, agatpm1c), greater than zero, the amount present is then modeled. As this variable is highly skewed, the amount of CAC present is transformed using a log transformation.

SAS 9.2 code example:

```
proc genmod data = b descending;
class idno race1;
where  agatpm1c ne 0;
model  log_transformed_CAC=age1c male bmi1c race1
     male  diabetes smoker ex_smoker  sbp1c dbp1c
   hdl1 ldl1 TRIG1STTN1C asa1c nsaid1c cox21c;
 repeated subject = idno/ type = unstr;
run;
```

The modeling of CAC is a good example of one of the more complicated continuous variables that can be encountered in CER.[Error! Bookmark not defined.] To properly model this association, two models were needed (and the second model required transformation of the exposure). Most comparative effectiveness projects will involve much simpler outcome variables, and the analyst should be careful to only include more complex models where there is an important scientific rationale.

Presentation

Often sensitivity analyses conducted for a specific CER study can simply be summarized in the text of the paper, especially if the number of scenarios is small.[17] In other cases, where a broad range of scenarios are tested,[2] it may be more informative to display analyses in tabular or graphical form.

Tabular

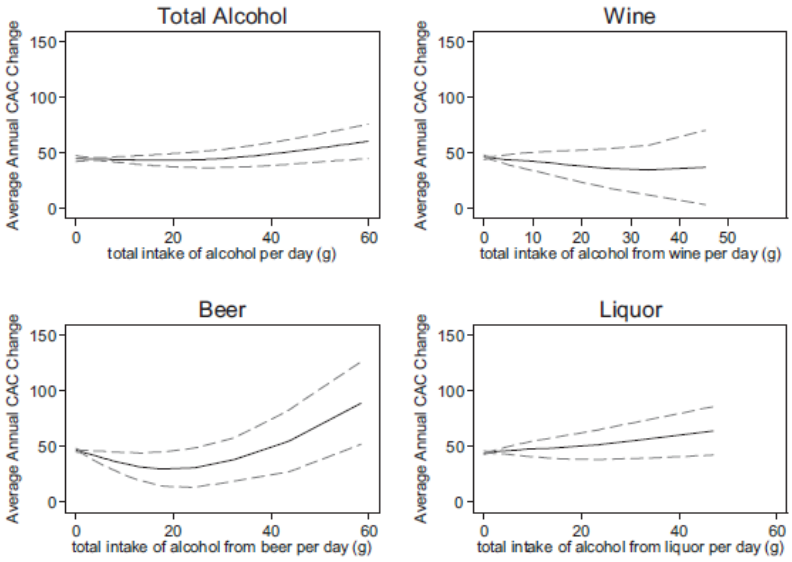
The classic approach to presenting sensitivity analysis results is in a table. There, the author can look at the results of different assumptions and/or population sub-groups. Tables are usually

preferred in cases where there is minimal information being presented, as they allow the reader to very precisely determine the influence of changes in assumptions on the reported associations. This is the approach used by Samy Suissa[14] to show differences in results based on different approaches to analyzing a cohort of lung cancer patients.

Graphical

One reason to use graphical methods is where the variable being modeled is, itself, a continuous variable, and presenting the full plot is more informative than forcing a categorization scheme on the data. One example, from Robyn McClelland and colleagues (Figure 11.1),[23] is a sensitivity analysis to see if the form in which alcohol is consumed changes its association with levels of CAC. The analyst, therefore, plots the association with total alcohol consumed overall and by type of alcohol (beer, wine, hard alcohol). Here, both the exposure and the outcome are continuous variables, and so it is much easier to present the results of the sensitivity analysis as a series of plots.

Figure 11.1. Smoothed plot of alcohol consumption versus annualized progression of CAC with 95% CIs



See McClelland RL, Bild DE, Burke GL, et al. Alcohol and coronary artery calcium prevalence, incidence, and progression: results from the Multi-Ethnic Study of Atherosclerosis (MESA). Am J Clin Nutr 2008 Dec;88(6):1593-601. This figure is copyrighted by the American Society for Nutrition and reprinted with permission.

Another reason for a graphical display is to present the conditions that a confounder would need to meet in order to be able to explain an association. As discussed, the strength of a confounder depends on its association with the exposure, the outcome, and its prevalence in the population. Using the standard spreadsheet discussed earlier,[20] these conditions can be represented as a plot. For example, Figure 11.2 presents a plot based on data from Psaty et al.[1,24]

Figure 11.2. Plot to assess the strength of unmeasured confounding necessary to explain an observed association

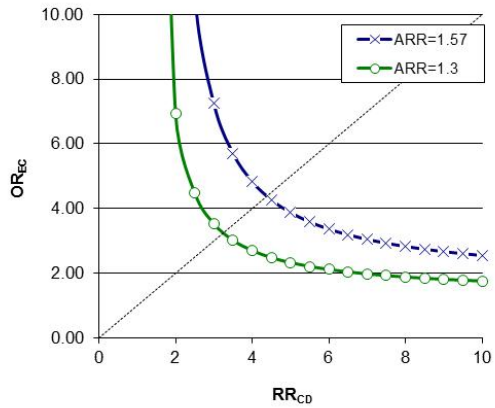


Figure 11.2 plots the combination of the odds ratio between the exposure and the confounder (OR_{EC}) and the relative risk between the confounder and the outcome (RR_{CD}) that would be required to explain an observed association between the exposure and the outcome by confounding alone. There are two levels of association considered (ARR=1.57 and ARR=1.3) and a separate line plotted for each. These sorts of displays can help illustrate the strength of unmeasured confounding that is required to explain observed associations, which can make the process of identifying possible candidate confounders easier (as one can reference other studies from other populations in order to assess the plausibility of the assumed strength of association). Other tools for sensitivity analysis are available, such as the one from Lash *et al.* (http://sites.google.com/site/biasanalysis/).[10]

Conclusion

While sensitivity analyses are important, it is necessary to balance the concise reporting of study results with the benefits of including of the results of numerous sensitivity analyses. In general, one should highlight sensitivity analyses that result in important changes or that show that an analysis is robust to changes in assumptions. Furthermore, one should ensure that the number of analyses presented is appropriate for illustrating how the model responds to these changes. For example, if looking at the sensitivity of results to changes in the exposure time window, consider looking at 30, 60, and 90 days instead of 15, 30, 45, 60, 75, 90, 105, and 120 days, unless the latter list directly illustrates an important property of the statistical model. The decision as to what are the most important sensitivity analyses to run will always be inherently specific to the problem under study. For example, a comparative effectiveness study of two devices might not be amenable to variations in exposure window definitions, but might be a perfect case for a physician preference instrumental variable. This chapter highlights the most common elements for consideration in sensitivity analysis, but some degree of judgment as to the prioritization of these analyses for presentation is required. Still as a general guideline, the analyst should be able to answer three questions:

- Is the association robust to changes in exposure definition, outcome definition, and the functional form of these variables?

- How strong would an unmeasured confounder have to be to explain the magnitude of the difference between two treatments?
- Does the choice of statistical method influence the directionality or strength of the association?

A plan for including some key sensitivity analysis in developing study protocols and analysis plans should be formed with a clear awareness of the limitations of the data and the nature of the problem. The plan should be able to answer these three basic questions and should be a key feature of any comparative effectiveness analysis. The use of sensitivity analysis to examine the underlying assumptions in the analysis process will build confidence as to the robustness of associations to assumptions and be a crucial component of grading the strength of evidence provided by a study.

DRAFT

Checklist: Guidance and Key Considerations for Sensitivity Analyses in an Observational CER protocol or proposal

Guidance	Key Considerations	Check
Propose and describe planned sensitivity analyses	- Consider the effect of changing exposure, outcome, confounder, or covariate definitions or classifications - Assess expected impact of unmeasured confounders on key measures of association	☐
Describe important subpopulations in which measures of effect will be assessed for homogeneity	- Consider pediatric, racial/ethnic subgroups, patients with complex disease states - Consider inclusion of AHRQ Priority Populations (http://www.ahrq.gov/populations/)	☐
State modeling assumptions and how they will be tested		☐
Indicate whether the study will be replicated in other databases, if available and feasible		☐

References

[1] Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 2006;15(5):291-303.
[2] McCandless LC, Gustafson P, Levy A. Bayesian sensitivity analysis for unmeasured confounding in observational studies. Stat Med 2007;26(11):2331-47.
[3] Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003;158:915-20.
[4] Rietbrock S, Plumb JM, Gallagher AM, van Staa TP.How effective are dose-adjusted warfarin and aspirin for the prevention of stroke in patients with chronic atrial fibrillation? An analysis of the UK General Practice Research Database. Thromb Haemost 2009;101(3):527-34.
[5] Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf 2010;19:858-68.
[6] Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, Manson JE, Robins JM.Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008 Nov;19(6):766-79.
[7] Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research.Clin Trials. 2011 doi: 10.1177/1740774511420743.
[8] Suissa S, Blais L, Spitzer WO, Cusson J, Lewis M, Heinemann L. First-time use of newer oral contraceptives and the risk of venous thromboembolism. Contraception 1997;56(3):141-6.
[9] Stürmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. Med Care 2007;45(10 Supl 2):S158-65.
[10] Lash TL, Fox MP, Fink AK. Applying Quantitative Bias Analysis to Epidemiologic Data. Springer, New York 2009.
[11] Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects--advantages and disadvantages. Nat Clin Pract Rheumatol 2007;3(12):725-32.
[12] Huybrechts KF, Brookhart MA, Rothman KJ, Silliman RA, Gerhard T, Crystal S, Schneeweiss S. Comparison of different approaches to confounding adjustment in a study on the association of antipsychotic medication with mortality in older nursing home patients. Am J Epidemiol 2011;174(9):1089-99.
[13] Cherubini A, Oristrell J, Pla X, Ruggiero C, Ferretti R, Diestre G, Clarfield AM, Crome P, Hertogh C, Lesauskaite V, Prada GI, Szczerbinska K, Topinkova E, Sinclair-Cohen J, Edbrooke D, Mills GH. The persistent exclusion of older patients from ongoing clinical trials regarding heart failure. Arch Intern Med. 2011 Mar 28;171(6):550-6.
[14] Suissa S. Immortal Time Bias in Pharmacoepidemiology. Am J Epidemiol 2008;167(4):492-9.
[15] Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 2009 Jul;20(4):512-22.
[16] Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 2010;19(6):537-54.
[17] Ionescu-Ittu R, Delaney JA, Abrahamowicz M. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. Pharmacoepidemiol Drug Saf. 2009 Jul;18(7):562-71.

[18] Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008;358(8):771-83.
[19] Delaney JA, Suissa S. The case-crossover study design in pharmacoepidemiology. Stat Methods Med Res 2009;18(1):53-65.
[20] Kronmal RA, McClelland RL, Detrano R, Shea S, Lima JA, Cushman M, Bild DE, Burke GL. Risk factors for the progression of coronary artery calcification in asymptomatic subjects: results from the Multi-Ethnic Study of Atherosclerosis (MESA). Circulation 2007;115(21):2722-30.
[21] Division of Pharmacoepidemiology & Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School. Downloads. Retrieved January 3rd, 2012 from http://www.drugepi.org/dope-downloads/.
[22] Walker AM. Observation and inference. An introduction to the methods of epidemiology. Newton Lower Falls, MA: ERI,1991.
[23] McClelland RL, Bild DE, Burke GL, Mukamal KJ, Lima JA, Kronmal RA; Multi-Ethnic Study of Atherosclerosis. Alcohol and coronary artery calcium prevalence, incidence, and progression: results from the Multi-Ethnic Study of Atherosclerosis (MESA). Am J Clin Nutr. 2008 Dec;88(6):1593-601.
[24] Psaty BM, Koepsell TD, Lin D, Weiss NS, Siscovick DS, Rosendaal FR, Pahor M, Furberg CD. Assessment and control for confounding by indication in observational studies. J Am Geriatri Soc 1999;47:749-54.