“Ten Commandments” for Conducting Comparative Effectiveness Research Using “Real-World Data”

Richard J. Willke, PhD, and C. Daniel Mullins, PhD

The use of “real-world data” (RWD), defined as “data used for decision-making that are not collected in conventional RCTs” (randomized controlled trials)¹ to inform comparative effectiveness research (CER) questions holds tremendous promise, which can be realized only if such research is conducted by strictly—religiously, one might say—following good research practices.^{2,3,4} The well-recognized potential for biases associated with analysis of nonrandomized data, as well as the increasing accessibility of these data and their potential for being data-mined, might lead some to view them as “forbidden fruit” for informing medical decisions.^{5,6} In fact, some may argue that RWD CER results based on nonrandomized data *a priori* compromises their credibility. Others may argue that clinical trials that target only regulators rather than post-regulatory decision makers, including patients, consumers, payers, prescribers, and policy makers are similarly, albeit differently, flawed because they are less informative for medical decision making than pragmatic clinical trials that address patient, prescriber, and payer concerns. In both randomized trials and studies using data with nonrandom assignment, the virtues of RWD CER results are more likely to be valued by appropriately skeptical audiences if decision makers are confident that the work has been conducted and reported with a dedication to high standards.

In this spirit of devotion to good research practices for CER using RWD, we offer “ten commandments” for conducting and reporting CER based on analysis of RWD, without any claim of having received them from on high. The purpose of this article is to provide the beginning-to-intermediate practitioner or decision maker with a concise list of practices that are crucial to the proper execution of this kind of work. It is not meant to replace the growing literature which, in many cases, more extensively reviews important technical aspects of RWD analysis, and we strongly recommend that readers also review other guidance documents and Task Force reports, such as those published by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). However, we believe there is merit in a brief overview of some key tenets of the RWD research process, from planning, to analysis, to reporting, that combine general good research practices with considerations specifically relevant to CER with RWD. Before beginning, we strongly urge those who conduct RWD studies to involve those who are part of the RWD data generation and decision-making processes when designing CER studies. This will maximize the usefulness of the RWD CER results.

I. Design your study to address the 3 central pragmatic features of CER, all oriented to informing a specific treatment

choice: active comparators; relevant patient populations; and outcomes that are meaningful to patients, prescribers, payers and policy makers.

CER is intended to improve the evidence base for making decisions that impact the health of “real world” patients. Thus, CER studies should make comparisons—directly or indirectly—of the drug or medical technology being studied to other medical technologies that are commonly used or recommended to treat the targeted indication. The comparators should be selected from among those most frequently prescribed as well as those recommended in clinical practice guidelines. CER must be pragmatic in nature, reflecting a reasonable cross section of patients who are likely candidates for the comparators being studied.⁷ Study outcomes, including the measurement, frequency, and timing of reporting outcomes, must be meaningful to patients and their providers, as well as payers and policy makers who affect access to drugs and other medical technologies. In order to be meaningful, outcomes must be relevant and important to patients; however, in a CER study, it also must be the case that outcomes vary across comparators and patients.⁸ That is, there must be a plausible causal relationship between the treatment and the meaningful outcomes and a recognition that the relationship may vary across subgroups.

As with all components of CER study design, analysis, and interpretation, stakeholder engagement can help to assure that the study is appropriately designed to be maximally relevant and informative for decision making. When CER is conducted with a particular payer or subgroup of patients in mind, the comparators, patient population, and outcomes should reflect that perspective.

II. Develop your research question such that all benefits and harms relevant to the treatment decision for the product relative to the comparator are considered. The research question must be well-defined *a priori* and targeted to provide a clear answer for a specific audience. Choose a research design (e.g., case-control, cohort) and a corresponding data-set (right population, right variables, large enough sample) that are likely to be able to answer and are suitable for your research question).

Both the blessing and the curse of large RWD sets are the many research questions that can be addressed with them, making them ideal for exploratory data analysis. However, when the goal is to present evidence on a question as outlined in Commandment I, especially for decision-making purposes, one’s work must be free of any suspicion that the bulls-eye was painted around the arrow. Just as a conventional RCT starts with a research question, with the subsequent protocol

and data collection designed specifically and parsimoniously to answer a pre-specified question, a CER study using RWD must start with a clear objective, which is usually best framed as a research question. Sometimes a research question begins as a very specific one, either to replicate or extend previous research. More often it begins rather broadly (e.g., "Does medication X result in better outcomes than medication Y in treatment of Z?"). If a broad CER question is being posed, then all benefits and harms of both products relevant to the treatment decision should be included in the analysis.

Before the analysis begins, a number of more specific conditions need to be imposed to clarify the question—which patients, with which characteristics, over what timeframe, under what definition of medication use, etc.? Those conditions should be based on what questions prior studies explored or left unanswered, or which specific coverage or treatment decisions require better information. Before a specific research question can be finalized, potential data sources should be reviewed for their feasibility (e.g., presence of the right variables, enough patients, and proper time frame) for answering the research question; even a very good data source may additionally delimit the research question. Finally, the specific research question, as well as the nature of disease and its treatment process, the prevalence of the outcomes of interest, the data source, and other factors will help determine the most appropriate research design (e.g., case-control, cohort, case-crossover, etc.).⁹ In the end, the research question should lead to an analytic framework that directly can test the hypothesis present in the question in a scientifically rigorous and informative manner. If the research question and relevant hypothesis cannot be tested satisfactorily with the RWD available, the research should likely not be done.

III. Investigate your data sources to understand the "real-world" process by which the data are generated. Describe the limitations of the data, as well as how patients are selected into or exposed to treatment, and when appropriate, describe potential concerns (e.g., classification bias, immortal time bias, adherence concerns, etc.) and how they are addressed.

Whoever said "what you don't know can't hurt you" never worked with RWD. Data are an inherently imperfect representation of the underlying characteristics they are meant to measure, even when collected following a strict protocol. Considering the highly variable conditions under which RWD are collected, recorded, transmitted, merged, etc., it's best to ask yourself, and possibly others, questions about any datum important to your study. Are data complete and, if not, are data missing at random or is there a systematic bias in under-reporting that could impact the results of the study? Why are different diagnosis codes for a condition used? Do those codes vary by location or other factors, and if so, for transparent

reasons? Are the patients or physicians not representative of typical practice in some way? Could their choices of treatments be limited by external factors, such as formulary restrictions or insurance provisions? What drugs used by patients may be missing from the data, and why? Incorrectly attributing exposure to treatment is called "classification bias."² Given test result data (e.g., blood pressure) under what conditions were those data collected, and why?

What can you do? Thoroughly review any underlying data manuals and/or questionnaires when they are available. While staying blinded to outcomes by treatment group, examine not only the descriptive statistics of key variables but also the distributions and lots of cross-tabulations. Consider consulting with a practicing physician, pharmacist, or a billing department employee to test your assumptions about your data. In addition, when constructing any outcome or control variables, be careful not to introduce any biases. For example, in a time-to-event analysis, introducing any time period during which the outcome could not have occurred will create "immortal time" bias.¹⁰ When categorizing patients based upon treatment, perform sensitivity analysis to see whether using different codes or time periods for exposure would affect how patients are categorized. When constructing total costs, be cognizant of systematic reasons why certain costs may be missing and exacerbate differences between treatment groups. In the end, it's never possible to find or adjust for all the imperfections in one's data, but doing the due diligence needed to be reasonably confident that the data are fit for the research task at hand is a fundamental responsibility of any empirical researcher. As in Commandment II, if the data are not deemed to be fit for the task, the research should not be continued; if the question is sufficiently important, a prospective study may be necessary.¹¹

IV. Write a full statistical analysis plan *a priori* that reflects current knowledge about comparator products and the evidence gap to be addressed; document any changes made along the way.

A pre-specified, well-written statistical analysis plan for a CER study provides benefits that are similar to those achieved by a pre-specified analysis plan for a conventional RCT. Having a roadmap provides a predetermined course for conducting the analysis and prevents deviations that otherwise could unintentionally change the validity or overall intent or direction of the study. It also avoids *post hoc* or selective reporting that tends to reduce the value and believability of results in the eyes of many decision makers. In fact, excessive *post hoc* analysis almost guarantees that certain results will appear to be statistically significant by chance rather than by true causation. Thus, pre-specified analysis plans enhance the credibility, efficiency, reliability, validity, and transparency of CER studies.

The statistical analysis plan should reflect a scientifically rigorous and clinically meaningful approach to answering the

study question. There should be specific aims and testable hypotheses that are directly related to the overall study objective and research question that are relevant for the comparator therapies being assessed. The analytic approach should be informed by what is known about the disease or condition being studied as well as the comparators being evaluated (see Commandment II). The statistical analysis plan should identify pre-specified subgroup analyses, specific codes (e.g., *International Classification of Diseases, Ninth Revision, Clinical Modification* [ICD-9-CM] and Current Procedural Terminology [CPT] codes) for inclusion/exclusion criteria, and the general approach for both descriptive and multivariable analyses (see also Commandments VI and VII).

As with conventional RCTs, there may be necessary deviations from the original statistical analysis plan because new evidence emerges from outside the trial or because of unexpected findings during the implementation of the analysis that require additional exploration. It also may be possible that the original statistical analysis plan failed to address a particular element appropriately. In all cases in which an amended analysis plan is required, it is important to report transparently not only what part of the statistical analysis plan was altered but why it was changed.

V. Carefully review univariate statistics for patient characteristics, outcomes, and control variables and how they differ across comparators. Investigate thoroughly the nature and degree of missing data (attrition, nonresponse, noncoverage, etc.) or miscoding, including anything that may affect treatment group identification.

In following Commandment III, you should have investigated some of these same issues in order to ensure that it was feasible to answer your research question with your data. Commandment V concerns the data analysis needed to inform not only yourself but also your audience about the nature of the data, its strengths and weaknesses, and its potential biases. The analysis begins with a thorough review of each relevant variable—outcome or control—and how it is distributed across comparison groups and across other relevant treatment subgroups. By identifying any fundamental imbalances, this descriptive analysis should inform and support any subsequent stratified or multivariate analyses. While this analysis cannot reasonably include "all possible" cross-tabulations, it should follow a logical process that ensures review of potentially important bivariate relationships, such as outcomes by disease severity across treatment groups.

A key aspect of this univariate review is attention to missing data. When control variables are missing, one should examine differences in outcomes across treatment groups for those with such variables present versus missing, in order to understand the biases that may be introduced by excluding observations with control variables missing. In cases where it appears that

control variables are missing at random, it may be feasible to impute them. Several techniques to handle missing data exist (e.g., listwise deletion, pairwise deletion, or multiple imputation), and the reader is encouraged to carefully consider the pros and cons of each method.^{12,13,14} Missing outcome variables or completely missing observations are generally more problematic, but methods are available to at least partly manage those problems.^{15,16,17,18,19,20,21} Sometimes missing data can lead to poor treatment group identification, called classification bias (see Commandment III). The key task at this stage of the analysis is to analyze and report the extent of the missing data as well as any information about why it occurred that can guide subsequent analysis.

VI. Control for observed confounders and other effect modifiers (explanatory variables) in a systematic and unbiased fashion and pay particular attention to how these may vary across comparator treatments; be wary of their correlation with the treatment variable. Choose 1 or more methods to address unobserved confounders (also known as selection bias); none is perfect and comparisons of different methods can be informative.

In RCTs, both effect modifiers (factors which affect outcome but not treatment choice) and confounders (factors affecting both outcome and treatment choice) are randomly, and in large trials, generally equally distributed between treatment groups, making explicit controls for these factors unnecessary to estimate unbiased average treatment effects. Nevertheless, a pre-specified multivariate analysis controlling for patient characteristics that affect treatment outcome can reduce residual variance, result in a smaller confidence interval on the treatment effect estimate, and using interactions, potentially identify treatment-effect heterogeneity.

Outside of RCTs, treatment groups are rarely balanced on observed characteristics and the potential for confounding of outcomes by unobserved factors is high. Physicians and patients commonly make choices about treatments based on factors that also affect treatment outcomes (e.g., patients who are more severely ill [in ways sometimes not observed] are often treated more aggressively, making the more aggressive treatment *a priori* biased towards having worse outcomes). Treatment effect estimates that don't both control for observed factors and consider unobserved factors are likely to be significantly biased. The literature on these issues is vast and distributed across statistical, econometric, epidemiological, psychological, and other disciplines. An overview of methods in this area, such as propensity score matching, stratification, instrumental variables, and others, as well as an extensive set of references, is found later in this supplement (Alemayehu et al., pages S22-S26).

Concerns around use of these methods can be grouped into 2 points. First, there are many choices of methods, including

selection of control variables, for a given problem, and each one may yield a different treatment-effect estimate. To avoid the temptation or appearance of picking a method *post hoc* that gives a "desirable" answer, the methods need to be clearly specified *a priori*. Second, one cannot know that a given method is going to give the "right" answer; none of the known methods can fully adjust for unobserved influences. Comparing the results of several methods, via a sensitivity analysis or simulation, given a sense of the strengths of each one, can provide insight into the robustness, or lack thereof, of conclusions around the comparative effectiveness estimates obtained.²²

VII. Choose a statistical technique and functional form for your estimation that is most appropriate to the outcomes of interest (time to event, linear regression, 2-part model, general estimating equations, etc.) across therapies as well as the relationship between treatment, confounders, and outcomes.

There frequently may be more than 1 analytic approach and multiple ways in which a regression equation can be specified to answer a particular CER question; however, there usually is one that is preferable based upon the study perspective, the conceptual design, or the data-generating process. While it may sometimes seem that no matter what you select, peer reviewers prefer an alternative statistical approach, it is important to remember that part of the responsibility of conducting a study is describing the pre-specified methods and defending why the specific statistical technique and functional form were selected. There is both a science and an art to conducting CER, and the best research balances the 2 considerations. The art of CER requires that the analytic approach is informed by clinical practice and patient decision making so that the regression results provide meaningful and interpretable output. The science of statistical analysis provides guidance for assuring that one can draw conclusions from the results because the statistical technique is appropriate and the functional form has been informed by model specification tests. It is equally important to pre-specify such alternative model specifications in the analysis plan and follow up the analysis with statistical testing of alternative functional forms. Specification testing provides critical information for determining whether there are interaction effects (e.g., whether the treatment effect varies by age or other observable patient characteristics), whether higher-order terms are required (e.g., whether variables are related in linear or nonlinear ways), and whether variables should be continuous or categorical. At the same time, it is always important to review the final regression approach and results for clinical plausibility.

VIII. Report univariate and multivariate results in an unbiased and complete fashion such that the benefits and risks of all comparators reflect "fair balance."

No study can answer all important questions, nor should one report every single data run, yet every study must provide objective and balanced reporting of the most relevant results regarding benefits and risks of all comparators included within the analysis. To achieve this balance, it is important to consider the viewpoints of decision makers, who are interested in comparisons of all clinically relevant benefits and potential side effects of treatments. This list should be informed by what is known or suspected about all treatment options included within the study. While all pre-specified outcomes should be provided in tabular form, it may be appropriate to highlight only those benefits and risks that are statistically different between the comparator treatments; however, in other cases, it may be important to comment on the fact that there is not a difference in key clinical outcomes. The reporting should include sufficient detail on the methods and results, including those from any alternative statistical approaches used, to provide the reader with a reasonably complete picture of the analyses performed; an online appendix can be useful for this purpose.

Objective reporting of outcomes requires that all benefits of all comparator treatments are given equal weight. Unfavorable outcomes should not be downplayed or "explained away." It is acceptable to translate the clinical importance of both positive and negative impacts of therapies on health so long as this, too, is done in with fair balance.

IX. Do not "over-interpret" results in the Discussion or Conclusion sections; remain objective in describing differences in outcomes across comparators.

The Discussion section should interpret CER results for key stakeholders and decision makers and place the study's results in context with prior knowledge and publications. Authors should comment on why comparative effectiveness results seem plausible, how the magnitudes of relative benefits and harms compare with those reported in prior studies, and whether observed differences are clinically and statistically significant. Although the Discussion may be somewhat subjective in nature, the interpretation of results should reflect an objective evaluation of what an unbiased individual could reasonably conclude from the study design and results. Authors should be careful to accurately reflect whether causal inference or correlation has been established and should avoid generalization of comparative benefits or harms beyond the study population and time frame.

Similar to regulators examining claims, payers and journal editors express strong criticism of manufactured-sponsored CER studies that appear partial in selecting which results are highlighted in the Discussion and Conclusion sections. Guidance on transparency in reporting can be found in the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement.²³ The focus of the entire

paper, including the Discussion and Conclusion sections, is driven by the research question (see Commandment II). Therefore, the importance of following all commandments simultaneously is essential.

X. Know and follow any external requirements (e.g., from ethics committees, federal or local governments), as well as any internal organizational protocols or SOPs for RWD study conduct and reporting.

Use of RWD data for CER studies is increasingly considered to impose 2 ethical obligations on the researcher—to use the data in sanctioned research and to report the results. Conditions for use of individual patient data are set by the owner of the data. In some cases, the research proposal must be reviewed by either the data owner or an ethics committee before the research can be carried out. Their intent is generally to protect patient confidentiality and ensure that the research is conducted along whatever lines were agreed to by the patients when their data were collected. Once the research is complete, it may be necessary to post results of safety-related or effectiveness-related outcomes in the spirit that it would be unethical to withhold potentially relevant information about treatments from the public. The state of Maine has required that RWD CER studies examining safety and effectiveness outcomes of drugs be posted in a way similar to posting of clinical trials.²⁴

Some institutions and companies have created their own standard operating procedures to provide both information and processes for employees to follow that help them comply with these obligations. For example, an institution may require that study protocols for both randomized trials and observational studies be posted on www.ClinicalTrials.gov and study results be posted in the ClinicalTrials.gov Results Database; we would encourage this practice even if it is not required. Researchers should ask data providers and their own institutions about such requirements before engaging in CER studies using RWD.

Summary

Decision makers want RWD studies and CER that provide meaningful evidence about the benefits and harms of alternative treatments. At the same time, they remain skeptical when RWD studies are not appropriately designed to answer relevant questions in a scientifically rigorous and transparent manner. While our proposed "ten commandments" cannot guarantee that studies are free from bias or other flaws—they can only address the devils you know—they nevertheless can serve as a useful checklist for improving the systematic use of principles that are aimed at achieving the goals of developing credible and germane CER studies using RWD.

Authors

RICHARD J. WILLKE, PhD, is Head, Global Health Economics & Outcomes Research, Global Market Access, Primary Care, Pfizer, Inc., New York, New York. C. DANIEL MULLINS, PhD, is Professor, Pharmaceutical Health Services Research Department, University of Maryland School of Pharmacy, Baltimore, Maryland.

DISCLOSURES

This supplement was funded by Pfizer, Inc. Willke is a Pfizer employee. Mullins reported financial and other relationships with Pfizer that include receipt of grants, consulting fees or honoraria, support for travel, consulting fees for participation in review activities such as data monitoring boards, payment for writing or reviewing the manuscript, advisory board membership, payment for lectures including service on speakers bureaus, and payment for development of educational presentations.

Willke and Mullins contributed equally to the concept and design, and writing and revision of the manuscript.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful comments of a number of colleagues, including Demissie Alemayehu, PhD; Jens Grueger, PhD; Frank Jen, PhD; Jack Mardekian, PhD; Andreas Pleil, PhD; and Robert J. Sanchez, PhD.

REFERENCES

1. Garrison LP, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force Report. *Value Health*. 2007;10(5):326-35. Available at: http://www.ispor.org/workpaper/RWD_TF/ISPORRealWorldDataTaskForceReport.pdf. Accessed September 23, 2011.
2. Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. 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 Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. *Value Health*. 2009;12(8):1053-61. Available at: <http://www.ispor.org/taskforces/documents/rdpartii.pdf>. Accessed September 24, 2011.
3. Helfand M, Tunis S, Whitlock EP, et al.; Methods Work Group of the National CTSA Strategic Goal Committee on Comparative Effectiveness Research. A CTSA agenda to advance methods for comparative effectiveness research. *Clin Transl Sci*. 2011;4(3):188-98.
4. Psaty BM, Siscovick DS. Minimizing bias due to confounding by indication in comparative effectiveness research: the importance of restriction. *JAMA*. 2010;304(8):897-98.
5. Austin PC, Platt RW. Survivor treatment bias, treatment selection bias, and propensity scores in observational research. *J Clin Epidemiol*. 2010;63(2):136-38.
6. Bosco JL, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *J Clin Epidemiol*. 2010;63(1):64-74. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2789188/?tool=pubmed>. Accessed September 24, 2011.
7. Horn SD, Gassaway J. Practice based evidence: incorporating clinical heterogeneity and patient-reported outcomes for comparative effectiveness research. *Med Care*. 2010;48(6 Suppl):S17-S22.

8. Wu AW, Snyder C, Clancy CM, Steinwachs DM. Adding the patient perspective to comparative effectiveness research. *Health Aff (Millwood)*. 2010;29(10):1863-71.
9. Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. *Value Health*. 2009;12(8):1044-52.
10. Suissa S. Immortal time bias in pharmacoepidemiology. *Am J Epidemiol*. 2008;167(4):492-99. Available at: <http://www.ispor.org/taskforces/documents/RDPartI.pdf>. Accessed September 23, 2011.
11. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*. 2000;21:121-45.
12. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Sons; 1987.
13. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. 1988;33(4):545-71.
14. Pickles A. Missing data, problems and solutions. In: Kimberly Kempf-Leonard, ed., *Encyclopedia of Social Measurement*. Amsterdam: Elsevier; 2005:689-94.
15. Hedeker D, Gibbons RD. Application of random effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*. 1997;2(1):64-78.
16. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley; 1987.
17. Briggs A, Clark T, Wolstenholme J, Clarke P. Missing...presumed at random: cost-analysis of incomplete data. *Health Econ*. 2003;12(5):377-92.
18. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8(1):3-15.
19. Horton NJ, Lipsitz SR. Multiple imputation in practice. *Am Statistician*. 2001;55(3):244-54.
20. Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics*. 1997;53(2):419-34.
21. Bang H, Tsiatis AA. Median regression with censored cost data. *Biometrics*. 2002;58(3):643-49.
22. Sturmer T, Schneeweiss S, Rothman K, Avorn J, Glynn R. Performance of propensity score calibration--a simulation study. *Am J Epidemiol*. 2007;165(10):1110-18. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1945235/?tool=pubmed>. Accessed September 23, 2011.
23. Vandembroucke JP, von Elm E, Altman DG, et al.; STROBE initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Intern Med*. 2007;147(8):W163-94. Available at: <http://www.annals.org/content/147/8/W-163.long>. Accessed September 23, 2011.
24. Maine Center for Disease Control and Prevention. *Clinical Trials*. Available at: http://www.maine.gov/dhhs/boh/clinical_trials.htm. Accessed July 10, 2011.