Comparative Effectiveness Research: What Kind of Studies Do We Need?

John Concato, MD, MS, MPH,*† Peter Peduzzi, PhD,*‡ Grant D. Huang, PhD,*¶
Timothy J. O’Leary, MD, PhD,†§ and Joel Kupersmith, MD¶

Abstract: Comparative effectiveness research (CER) is increasingly popular, yet discussions of its conduct and consequences often overlook the extensive history of comparing different therapeutic options in patient-oriented research. In particular, research in the Department of Veterans Affairs (VA) has included a decades-long focus on generating information that can enhance medical decision making and improve health outcomes. Categories of such research include multisite randomized controlled trials (conducted by the Cooperative Studies Program) and observational studies involving either primary or secondary data collection. As representative examples from cardiology, a landmark VA clinical trial published in the 1970s evaluated the benefits of coronary artery bypass grafting surgery among patients with angina; a VA trial initiated in the 1990s, and identified formally as CER, demonstrated that percutaneous coronary intervention is not superior to optimal medical therapy; and a database investigation using information from the VA electronic medical record system in the 2000s found that use of proton pump inhibitor medication is associated with the attenuation of the benefits of clopidogrel among patients hospitalized for acute coronary syndrome. A review of these (and other) selected projects, based on their type of study design, serves to highlight the strengths, limitations, and potential of CER.

Key Words: comparative effectiveness research, randomized controlled trials, observational studies, database research, meta-analysis, US veterans

(J Investig Med 2010;58: 764–769)

Comparative effectiveness research (CER) evaluates the relative effectiveness of different options for treating a specific medical condition in a selected population.1–4 Comparative effectiveness research is not a single entity; it can focus on patient-level or system-level elements of care, involving medications or procedures, with or without considering costs. Regardless of its scope, CER has the potential to generate information that can enhance medical decision making and improve health outcomes, but concerns exist regarding its conduct and consequences.5,6

Recent attention to CER has discussed mechanisms to implement findings into practice,7,8 potential impacts on personalized medicine,9 and disagreements regarding funding.5 Less attention has been given to strategies of the design and conduct of such research. Whereas discussions of nonscientific issues are ongoing, an internal look at the methods of CER is also warranted. In addition, selected examples from the Department of Veterans Affairs (VA) can illustrate one organization’s experience in conducting CER.

HISTORICAL OVERVIEW

Comparative effectiveness research can be considered a subset of patient-oriented research that examines available therapeutic options among particular patients to determine relevant health outcomes. As such, CER is a relatively recent term, but the idea of evaluating “real world” treatment options dates back to antiquity. As a centuries-old example, James Lind compared various treatments for scurvy in the mid-1700s10; providing citrus fruits led to resolution of symptoms, whereas vinegar, cider, and other remedies did not. (This example also demonstrates that valid associations can be identified before a full explanation of underlying mechanisms; the isolation of vitamin C and a biological understanding of scurvy did not occur until almost 200 years later.) In the 20th century, the so-called comparative effectiveness of 2 penicillin treatment schedules in pneumococcal infections of mice11 was evaluated in 1948, but the phrase “dosing study” might be applied to the same research today. In the same era, the first rigorous studies of the treatment of tuberculosis in humans involved the US Veterans Administration, Army, and Navy collaborating to evaluate streptomycin among veterans and soldiers whose condition was diagnosed after World War II.12,13 This observational study predated, and found results similar to, a randomized trial of streptomycin conducted in the United Kingdom by the Medical Research Council.13 (The latter study13 is acknowledged as the first randomized trial reported in modern clinical medicine.)

In the last half-century, terminology for research design has evolved, based specifically on the meaning and application of the words “efficacy” and “effectiveness.” Attributable in large part to the passage (in 1962) of the Kefauver-Harris Amendment14 to the Federal Food, Drug, and Cosmetic Act and developed in the wake of thalidomide causing phocomelia and other birth defects, this amendment required proof of efficacy as part of the drug approval process in the United States. The subsequent adoption of phase I to III (and later phase 0 and phase IV) schema for clinical trials emphasizes the evaluation of new drugs versus placebo rather than with existing agents. This system also places more emphasis on internal validity — based on “efficacy trials” that tend to have restrictive inclusion, exclusion, and treatment criteria — than on external validity derived from “effectiveness trials” that tend to have patient populations and (nonplacebo)
interventions more representative of clinical practice. Implementation of research results after the publication of trials (or other types of studies) is another relatively recent consideration. By any reasonable definition, numerous studies of comparative effectiveness have been done in recent decades, but usually without being identified as such. Interestingly, the formal term “comparative effectiveness research” has been popularized only in the past several years, related more to policy and economic considerations than to scientific issues themselves. Whether an article is classified as comparative effectiveness is variable, depending on the authors and the viewpoint of reviewers and journal editors. A recent review classified 689 studies as CER based on a search of clinicaltrials.gov, a database of clinical trials, and HSRProj, a database of health services research projects.

The visibility of CER was increased by an appropriation of $1.1 billion in the American Recovery and Reinvestment Act of 2009. Although this amount is less than 1% of the $150 billion directed toward health care overall, the support represents a major infusion of funding for CER. Even before this new initiative, however, CER had become a prominent component of research on health care. As the purposes and implications of CER are debated in medical and nonmedical venues, a major unanswered question is what kind of CER will be considered worthwhile? For example, in the current era of evidence-based medicine, a hierarchy of research designs might be promoted for CER that favors randomized trials. Conversely, the relatively short timeframe to spend “stimulus funds” encourages the use of secondary sources of data that can be analyzed promptly.

CATeGORIES OF RESEARCH DESIGN

From a methodological perspective, randomized controlled trials can address comparative effectiveness and are arguably underused for this purpose. Specifically, studies funded by the pharmaceutical industry represent most trials in the United States but are usually designed to evaluate efficacy and not effectiveness. As mentioned, randomized trials evaluating efficacy (or explanatory clinical trials) often compare an intervention to placebo among patients with little or no comorbid illness to address whether an intervention can work under ideal circumstances. Early-phase clinical trials and applications to the Food and Drug Administration for licensure, including studies conducted or supported directly by the pharmaceutical industry, tend to fall into this category.

Randomized trials evaluating effectiveness (or “practical clinical trials”) determine whether an intervention actually works under more representative, or real-world, conditions. These studies compare an intervention to a standard-of-care among patients representative of a target population. Randomized trials have often been criticized as lacking generalizability, but effectiveness trials are more likely to apply to general medical practice than efficacy trials. Specific methodological strategies, such as adaptive designs, represent novel features of trials that can improve their utility.

Random allocation promotes a balance of prognostic factors among groups being compared—precluding a major source of bias—but randomized trials are not always feasible, ethical, or necessary. Well-designed observational cohort studies and case-control studies collecting primary data constitute another approach to CER. These studies have enhanced generalizability in representative patient populations (especially compared with efficacy trials), yet can produce results similar to randomized trials when sufficient attention is given to sources of bias. In this context, an often-cited example of hormone replacement therapy and subsequent health outcomes has generated misconceptions regarding the limitations of observational studies. In reality, when differences in patient populations, treatment regimens, and duration of follow-up are taken into account, data from randomized and observational studies on this topic (and others) are quite consistent. So-called database research, involving information obtained for clinical or administrative purposes during health care delivery, is another (observational) method for evaluating comparative effectiveness. The generalizability of results can become unrivaled, given assessments of large and representative populations of health care recipients. Conversely, problems include the lack of randomization and concerns with data quality (such as “Are diagnoses listed in an administrative database accurate?”). Carefully validated analytic strategies are therefore critically important; comparison of results with randomized trials and other observational studies can help to enhance interpretation.

These categories of research design are not exhaustive, but they do account for common, primary strategies of CER. Secondary strategies include systematic reviews and statistical modeling. Rather than obtaining data as observations on individual participants, systematic reviews (including meta-analyses) use information from previously completed studies—but not without limitations. “Secondary” statistical modeling uses novel techniques to assess cause-effect associations, typically with a data set that had already been collected and analyzed. Although most patient-oriented research investigations use a statistical model per se, this type of statistical modeling goes beyond original hypotheses and planned calculations to explore associations based on statistical theory and corresponding assumptions—that may or may not be justified.

EXAMPLES FROM THE DEPARTMENT OF VETERANS AFFAIRS (VA)

The VA has conducted CER using various methodologies for many years. For example, the VA Cooperative Studies Program has been recognized for producing “numerous studies designed to support clinical and policy decisions.” The Cooperative Studies Program has generated more than 175 large-scale, multisite investigations, mostly as randomized trials of comparative effectiveness. Key elements of the Program include methodological, technical, and administrative personnel in a national infrastructure of specialized centers for conducting collaborative research. The Program also has established guidelines for the design, conduct, and analysis of its multisite studies.

Selected trials over 4 decades have shown (Table 1) that coronary artery bypass surgery (vs medical therapy) improves survival among patients with left main coronary artery disease, steroid therapy (added to usual care) does not reduce mortality from septic shock, watchful waiting (vs transurethral resection) is a safe alternative among men less bothered by symptoms from benign prostatic hyperplasia, and percutaneous coronary intervention (vs optimal medical therapy) does not reduce the risk of cardiovascular events or death in stable angina. These trials were selected with the main goal of highlighting both a long-standing “VA experience” and a broad scope of topics and methods, such as medical versus surgical interventions and usual care versus optimal therapy as a comparator. In conducting CER, and looking beyond the strengths or limitations of any individual study, the Cooperative Studies Program has been described by non-VA authors as “an important source of practical clinical trials in the United States.”

© 2010 The American Federation for Medical Research

765
Observational studies with primary data collection also contribute to CER in the VA, originating from the Cooperative Studies Program (via epidemiology centers), individual investigators, and other sources. An example is the National Surgical Quality Improvement Program (NSQIP) that systematically measures and seeks to improve risk-adjusted surgical outcomes using validated methods. The NSQIP\(^{35}\) is based in the VA Health Services Research and Development Service and was expanded to the private sector by the American College of Surgeons.\(^{36}\) In a collaborative NSQIP\(^{37}\) project, abnormally low or high preoperative hematocrit levels were found to be associated with increased 30-day mortality and cardiac events after noncardiac surgery among elderly patients (Table 1). These results inform CER by identifying a potentially modifiable risk factor for developing targeted treatments, with a goal of improving outcomes.

Health services research projects frequently use archived database information only, as in a recent evaluation\(^{38}\) of an adverse interaction between proton-pump inhibitor medications and clopidogrel after hospitalization for acute coronary syndrome (Table 1). This study is pertinent in merging and analyzing national-level VA data collected for other purposes, with careful attention given to possible sources of bias and with multiple sensitivity analyses. The Department of Veterans Affairs Health Services Research also supports the Quality Enhancement Research Initiative,\(^{39}\) designed to translate study findings into clinical practice via implementation research.\(^{31,38}\)

As an example of a systematic review (Table 1), a report\(^{40}\) evaluated the comparative effectiveness of percutaneous coronary interventions and coronary artery bypass graft surgery (CABG); the focus was not limited to VA-based studies. After a structured literature search, results from 23 randomized trials were identified, reviewed, and analyzed. The authors concluded that compared with percutaneous coronary interventions, CABG was more effective in relieving angina, and led to fewer revascularizations, but had a higher risk for procedural stroke; “survival to 10 years was similar” for both procedures.\(^{40}\)

As an example of “secondary” statistical modeling, data were obtained from a VA Cooperative Study\(^{41}\) and a similar trial conducted among nonveterans,\(^{42}\) evaluating naltrexone in the treatment of alcohol dependence (Table 1). Although the original trials were null in terms of not finding an improvement in outcomes, such as the time to the first day of drinking a threshold amount of alcohol, questions arose as to whether the intervention might have benefited certain participants. A follow-up study\(^{43}\) used longitudinal models with “fixed polynomial trends over time” and other statistical methods; the analyses generated estimates of the proportion of participants classified (on a given day) as abstinence, drinking sporadically, or drinking consistently. The authors reported that “naltrexone may have a clinically meaningful effect for alcohol-dependent patients” regarding consistent drinking.\(^{43}\)

These VA-based (and other) examples serve collectively to characterize a methodological spectrum of CER. Within the Veterans Health Administration, such research has changed health policy and influenced clinical care,\(^{44}\) as with recommendations and treatment patterns for anemia after a study comparing subcutaneous and intravenous administration of erythropoietin.\(^{45,46}\) Despite such examples to support an argument that research can and does directly change practice patterns, a counterargument would include the previously cited example\(^{39}\) of a randomized trial evaluating CABG surgery for patients with angina. Approximately 10 years after publication of the main report, questions still lingered as to whether the study was valid and should influence the care of patients;\(^{47}\) the answers eventually emerged as “yes” and “yes,” respectively.\(^{48}\) Ultimately, the complex issue of how to change physicians’ behavior is beyond the scope of this report, but conducting and understanding rigorous and relevant CER is an important component of the process.

A related issue involves the patient population receiving health care at VA medical facilities, representing the source population for VA-based research. For example, although the proportion of women in the VA is increasing, their representation

| TABLE 1. Representative Examples of CER Conducted by VA, Organized According to Type of Study Design |
|-----------------------------------------------|------------------------------------------------------------------------------------------|
| Study Design and Clinical Context | Information to Improve Patient Care (Year) |
| Randomized controlled trials: | Coronary artery bypass grafting surgery improves survival compared with medical therapy (1976). |
| Patients with angina pectoris found to have a significant lesion in the left main coronary artery | Steroid therapy added to usual care is not effective in reducing mortality among patients with septic shock (1987). |
| Patients with clinical signs of systemic sepsis receiving intensive care | Watchful waiting is a safe alternative to surgery among men less bothered by urinary symptoms (1995). |
| Ambulatory men with symptoms of benign prostatic hyperplasia | Percutaneous coronary intervention is not superior to optimal medical therapy (2007). |
| Patients with evidence of myocardial ischemia and stable coronary artery disease | Abnormally low or high preoperative hematocrit levels increase mortality after surgery (2007). |
| Observational studies involving primary data collection: | Proton pump inhibitor medications are associated with attenuation of benefits of clopidogrel (2009). |
| Older patients undergoing major noncardiac surgery | Coronary artery bypass surgery improves angina and decreases revascularizations versus percutaneous intervention, but with higher risk of stroke and similar survival (2007). |
| Database (observational) investigations of archival information: | Naltrexone may reduce levels of “consistent” drinking even if time to first drink is not improved (2007). |
| Patients after discharge from a hospital admission for acute coronary syndrome | Permanent corrective surgery improves survival compared to medical therapy (1976). |
| Systematic reviews (including meta-analyses): | Percutaneous coronary intervention is not superior to optimal medical therapy (2007). |
| Patients with coronary artery disease considered eligible for coronary revascularization | Naltrexone may reduce levels of “consistent” drinking even if time to first drink is not improved (2007). |
| Secondary statistical modeling: | Permanent corrective surgery improves survival compared to medical therapy (1976). |
| Ambulatory patients receiving treatment for alcohol dependence | Permanent corrective surgery improves survival compared to medical therapy (1976). |
still falls far short of the one-to-one balance in the community. Similarly, veterans who receive VA health care tend to come from lower socioeconomic strata, often linked to private health insurance being less available or affordable to them. Thus, a question of generalizability (external validity) does exist when VA studies are considered, yet this concern is sometimes extrapolated inappropriately to cast doubt on the internal validity of a study, without giving due consideration to the methods involved. In other situations, this concern is not applicable—for example, when VA studies are done in collaboration with the National Institutes of Health or as international collaborations.

Overall, if examined as a body of knowledge, VA studies in CER have been considered reliable and relevant in improving care delivered to the general (nonveteran) population. In addition, an informal review (data not shown) indicates that VA research has contributed to a sizeable majority of the 100 categories listed in the Institute of Medicine’s recent report on national priorities for CER.

GENERAL CONSIDERATIONS

The VA is distinctive in supporting virtually all of the types of research design mentioned in this report and in promoting implementation of research results. Other organizations and agencies might embrace a hierarchy of research design so that randomized trials and meta-analyses are emphasized as the best forms of evidence. (Under such circumstances, the VA Co-operative Studies Program should be considered an important repository of randomized studies.) At the same time, however, any approach that imposes a research design hierarchy should be understood as placing a “filter” between the generation of evidence and its interpretation, bypassing the question of whether such filtering is appropriate or how well it is done. This problem can be subtle, as when “definitive randomized controlled trials” versus “weaker forms of evidence, such as nested case-control or well-designed cohort studies” are presented as self-evident truths in discussions of CER. Conversely, other organizations and agencies might use only secondary sources of information, perhaps underestimating the corresponding challenges, for reasons including database research can be done more readily.

A major take-home message of this report is that types of study design are an important feature of CER, but hierarchies of research design should not become a distraction in the process of scientific inquiry. A concurring opinion was articulated as: “This is not a plea to abandon RCTs and replace them with observational studies. […] Rather, it is a plea to investigators to continue to develop and improve their methods; to decision makers to avoid adopting entrenched positions about the nature of evidence; and for both to accept that the interpretation of evidence requires judgment.” A similar viewpoint appeared in a report on CER, with statements including “traditional ‘hierarchies of evidence’ are overly simplistic and should not necessarily guide the implementation of CER”; “a range of methods are important in CER”; and “the right approach to a given CER study depends on the circumstances.” (Of note, a recent descriptive report suggests that randomized trials constitute most of the published CER studies.)

Another take-home message involves the attributes needed for an organization to conduct CER. The VA is certainly not alone in its capabilities and commitment to CER, and it also is not the largest health research organization in the United States. Yet, the unique position of VA research in the national health care environment includes at least 4 elements that have sustained CER: a comprehensive and relatively stable infrastructure, support for methodological advances, research ideas arising from clinical and administrative questions, and a commitment to excellence in research on clinical care. Studies of comparative effectiveness in the VA are also facilitated by an electronic medical record system, a uniform reimbursement mechanism, and a national health care delivery network. In particular, few national-level organizations have a similar ability to integrate clinical, laboratory, pharmacy, service use, and outcome data in one health care context.

The VA also serves an example of an organization that can collaborate with the private sector and yet avoid common conflicts of interest, using a philosophy of “science comes first.” As exemplified by the Cooperative Studies Program, strict policies govern the (intramural) design, conduct, interpretation, and regulatory aspects of a study, whereas unrestricted research grants (or in-kind support, such as medications) from industry are acceptable. Cooperative Research and Development Agreements (CRADAs), authorized under the Federal Technology Transfer Act, represent one strategy used to safeguard the interests of the VA and veterans while being sensitive to the perspective of industry regarding commercially valuable information. Research conducted under a Cooperative Research and Development Agreement concurs with federal statutes, regulations, and policies and, at the same time, protects background inventions, trade secrets, and confidential information. A representative statement would mention that “[the industry sponsor] had no role in the design of the study, collection, and analysis of data, or preparation and approval of the manuscript.” Given the paucity of incentives for the pharmaceutical or medical device industry to conduct CER on their own, this category of research represents an important opportunity to advance knowledge. Finally, details pertaining to cost-effectiveness analysis are beyond the scope of this paper, but a decision not to focus on costs in studies of CER is related more to policy than to science. Interestingly, the integrated nature of the VA health care system is a potential advantage in terms of the ability to assess the costs of treatment options accurately on a head-to-head basis.

CONCLUSIONS

If a national-level entity dedicated to CER is created, either as a public or private (or hybrid) agency, the culture of such an organization would likely influence the selection of types of study design, but the methodological issues raised in this report should be considered. Rather than adhering to a hierarchy of research design, comparative effectiveness research should be conducted and evaluated using sound clinical, epidemiological, and statistical judgment. Beyond these attributes of CER, and in the broader context of health care reform, careful attention should be given to methods of patient-oriented research. Although not unique to CER, important steps include identifying an important research question; developing a cogent study design among feasible options; collecting relevant, high-quality data; and performing appropriate statistical analyses. Ultimately, precise ratings of interventions based on CER are arguably neither feasible nor desirable, and the source of CER is less pertinent than creating and disseminating rigorous research. The utility of a particular CER study should be judged in a peer-reviewed, scientific context and applied to individual patients by competent and caring clinicians.

REFERENCES


