Comparative effectiveness research: what is it and why do we need it in nephrology?

Tara I. Chang and Wolfgang C. Winkelmayer

Division of Nephrology, Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, USA

Abstract
The USA leads other industrialized countries in health care spending but lags behind in terms of health outcomes. There has been growing interest in comparative effectiveness research (CER) as a means to identify best practices to create a more efficient and effective health care system. Two key concepts of CER are that it should (i) compare two or more alternative tests, therapies or procedures and (ii) be conducted in persons, clinical settings and conditions that are representative of the real world. The goal of CER is to provide evidence for clinicians, patients, policy makers and others to make informed decisions that will ultimately improve the overall health of specific subgroups and of the population as a whole. In this narrative review, we first describe the strengths and limitations of various types of studies that constitute CER, including randomized clinical trials, observational studies and systematic reviews, providing examples from the nephrology literature. Because of the concerns regarding confounding in observational CER, we also provide an overview of methods to reduce confounding in these types of studies. Finally, we will discuss why CER pertaining to kidney disease care needs to be a top priority in order to move our field from a largely opinion-based specialty to an evidence-based specialty.

Keywords: comparative effectiveness; epidemiology; methods; outcomes; review

Introduction
Over the past 20 years, the USA has held the dubious distinction of being the country with the highest health care spending (as a percentage of gross domestic product), but one of the lowest overall life expectancies compared to other industrialized nations [1]. In addition, numerous studies have demonstrated considerable geographic variation in health care utilization, costs and outcomes [2, 3]. With health care spending continuing to skyrocket, especially in light of difficult economic times, the US federal government recognized the need to identify best practices in order to create a more efficient and effective health care system. In order to ameliorate the immediate economic consequences of the acute financial crisis of 2008, the 111th Congress passed the American Recovery and Reinvestment Act of 2009 [4]. As part of this enormous stimulus package of over US$ 700 billion, the US government allocated US$ 1.1 billion for comparative effectiveness research (CER) to help identify effective therapies for more informed medical decision making, thrusting CER into the national spotlight. In 2010, the Patient-Centered Outcomes Research Institute was created and tasked with conducting high-quality CER to help provide the necessary evidence for patients and providers to make informed treatment decisions [5]. Given the growing interest in CER in the USA and beyond, this review will provide an overview of what constitutes CER, how CER is performed and why we, as specialists in nephrology, need CER to inform our clinical practice.

What is CER?
Several definitions of CER have been developed, with the most relevant ones being from the US Federal Coordinating Council [6] and the Institute of Medicine [7] (Table 1). What is common to these definitions is that CER is not solely a comparison of one drug with another but meant to be extended to different clinical strategies that may involve medical or non-medical preventive strategies, diagnostic testing (single tests or several in sequence or combination) and different treatments, including medications, devices, surgeries or rehabilitative techniques. The focus is on ‘real-world’ settings as opposed to ‘laboratory’ settings as is the case in trials of highly pre-screened and selected patient populations. In addition, the results need to be useful for patients and providers in their shared decision making. Finally, the results should be broadly generalizable to the overall population and to specific subpopulations, especially ones that are often excluded or marginalized from trials: older patients, children, women, racial, ethnic or socioeconomic minorities, and patients with complex comorbid conditions such as chronic kidney disease (CKD).

Methods for CER
To achieve these goals, different types of research studies can constitute CER, including randomized clinical trials...
(RCTs), observational studies and systematic reviews and meta-analyses (Figure 1). Each type of study design, with its particular strengths and limitations, has an important and complementary role in CER.

**Randomized clinical trials**

RCTs can be divided into two general types: the explanatory RCT and the pragmatic RCT, though most RCTs fall along a continuum rather than being entirely of one type or the other [8]. RCTs are generally of the explanatory type, with the goal of answering the following question: ‘can this therapy work under ideal conditions?’ As such, the explanatory RCT employs fairly stringent inclusion and exclusion criteria to ensure adherence to the protocol, to reduce costs of conducting the study or for other valid and practical reasons. However, by limiting the diversity of enrolled participants as well as being usually limited to more specialized (academic or non-academic) research practice settings, the study may not reflect real-world clinical practice, one of the central concepts in CER. The issue of real-world applicability is especially relevant to the nephrology community, as patients with CKD are often excluded from RCTs. For example, of 153 major RCTs studying cardiovascular therapies, over half excluded CKD patients [9].

In contrast to the explanatory RCT, the goal of the pragmatic RCT is to answer the question: ‘does this therapy work under usual conditions?’ To do so, pragmatic RCTs have broader entry criteria to ensure greater patient diversity. Additionally, rather than having protocol-driven, tightly controlled follow-up visits, participants in the pragmatic RCT are followed in a more usual care fashion [10]. Consume with the real-world intention, patients may exhibit their usual tendencies to adhere to any interventions or not. As an example of a hypothetical pragmatic RCT, patients with CKD could be randomized to receive free passes for exercise classes to study the impact of physical activity on any number of clinical outcomes. Patients in the intervention arm may not open the study letter, choose the intervention arm may not open the study letter, choose to ignore it, or may participate in the exercise classes with varying persistence, while patients who do not receive the passes may not even know that they constitute the control group. All participants would be passively followed for the outcomes of interest through routine care as registered in electronic health records of insurers or health care systems.

Another way to conduct the pragmatic RCT is to use a cluster randomization technique, which randomizes groups rather than individual participants [11]. The patients may not know that they are participating in a trial and informed consent is usually not required. For example, a recent cluster RCT randomly assigned group practices to receive either an enhanced laboratory prompt containing specific guideline-based management recommendations when a patient’s estimated glomerular filtration rate (eGFR) was <60 mL/min/1.73 m² or to receive the standard laboratory reporting of the eGFR [12]. There were very few exclusion criteria, and because outcomes were ascertained through pharmacy claims, no additional trial-associated study visits, which might have influenced outcomes (in this case, use of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers), were needed. One drawback of cluster randomization is that it requires a larger number of study participants than would an RCT with individual-level randomization to achieve adequate statistical power [11].

Many consider RCTs to be the gold standard of CER, as the random allocation of the treatment of interest should in theory result in an even distribution of measured and unmeasured confounders across the treatment groups. Indeed, the randomization process is a major strength over non-randomized evaluation of interventions seen in observational studies. However, RCTs also have several important limitations. First, not all important clinical questions can be answered by an RCT, either due to ethics, cost or other issues of feasibility. For example, while observational studies have shown a higher risk of more rapid progression of CKD associated with tobacco smoking [13], a definitive RCT with participants randomly assigned to a smoking strategy would...
have great difficulty passing an institutional review board. Slowly-progressing conditions, such as development of end-stage renal disease (ESRD), are also challenging to study as an outcome in RCTs as the costs and resources that would be required to follow participants for years or even decades would be prohibitive. Because of these limitations, RCTs examining CKD progression have often needed to use surrogate end points (e.g. changes to serum creatinine, reduction of proteinuria) that are of questionable validity [14].

Additionally, the majority of RCTs are performed for regulatory approval of drugs or devices, with priorities set by industry rather than patients or clinicians, and these often favor comparison with placebo rather than active treatment. ‘True’ CER usually compares two active comparable therapies or strategies, although in special cases a ‘do nothing’ strategy may also constitute an important comparator as is the case in the hypothetical exercise trial above. For example, there was no incentive for industry to conduct a definitive RCT comparing the safety of rosiglitazone and pioglitazone. Large database studies provided evidence eventually leading to decisive action by regulatory agencies to ban or restrict the availability of rosiglitazone in 2010 [15–17].

Finally, it is not safe to assume that just because a study is an RCT that its results are valid. There are important metrics that should be assessed when evaluating the quality of an RCT and whether its results should be used in decision making. Several quality scales have been developed. Two of the most commonly used are the CONSORT [18] statement, endorsed by several leading scientific journals and used to improve reporting of the primary results or design of an RCT, and the Jadad Scale [19], often used to judge the quality of a previously reported RCT under consideration for inclusion in a systematic review or meta-analysis (Table 2).

### Observational studies

Observational studies can address many of the limitations of RCTs. Because observational studies often leverage data originating from clinical practice, they better reflect real-world conditions, a central concept of CER. For example, the United States Renal Data System (USRDS) [34] compiles information on essentially every patient undergoing chronic dialysis or recipients of a kidney transplant in the USA; there are no inclusion or exclusion criteria. Moreover, billing claims data are available for a large subset of these patients, allowing ascertainment of hospitalizations, laboratory and other diagnostic tests, procedures, outpatient physician visits, nursing home stays and other important patient-centered issues. With the increasing use of electronic health records in integrated health systems such as Kaiser Permanente or dialysis providers such as several subsidiaries of Fresenius Medical Care AG, large amounts of detailed real-world patient data are more accessible than ever before.

In addition, observational studies can examine relationships that would be difficult or impossible to assess with an RCT. For example, while RCTs usually compare a single intervention to another, observational studies can examine multiple treatment comparisons [35, 36]. Observational studies can study adverse drug effects or toxic exposures as well as slowly developing conditions like ESRD. Observational studies can also better examine

---

**Table 2.** Selected resources used to assess the quality of different CER study types

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Published examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized Clinical Trials</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSORT (consolidated standards of reporting trials) [18]</td>
<td>25-item checklist and flow diagram</td>
<td>TREAT study [20]</td>
</tr>
<tr>
<td>Endorsed by leading journals</td>
<td>Developed in 1996 for pain/analgesic RCTs</td>
<td>Sharif et al. JASN 2011 [22]</td>
</tr>
<tr>
<td></td>
<td>Scores range from 0 (lowest) to 5 (highest)</td>
<td>Sandhu et al. JASN 2006 [23]</td>
</tr>
<tr>
<td>Observational Studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STROBE statement (strengthening the reporting of observational studies in epidemiology) [24]</td>
<td>Several checklists for cohort, case-control and cross-sectional studies</td>
<td>Molnar et al. JASN 2011 [25]</td>
</tr>
<tr>
<td><a href="http://www.strobe-statement.org/">http://www.strobe-statement.org/</a></td>
<td>Endorsed by leading journals</td>
<td></td>
</tr>
<tr>
<td>GRACE principles (Good ReseArch for Comparative Effectiveness) [27]</td>
<td>Broadly applicable</td>
<td>Tentori et al. NDT 2012 [26]</td>
</tr>
<tr>
<td><a href="http://www.graceprinciples.org/">http://www.graceprinciples.org/</a></td>
<td>High-level guidance regarding study plan, conduct, analysis and validity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Focused specifically on CER</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GRACE checklist currently undergoing validation/testing</td>
<td></td>
</tr>
<tr>
<td>ENCePP guide on methodological standards in pharmacoepidemiology (European Network of centers for pharmacoepidemiology and pharmacovigilance) [28]</td>
<td>Focus on medications</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.enc%D0%B5%D0%BFp.eu/index.html">http://www.encепp.eu/index.html</a></td>
<td>Overview of available methods for high-quality pharmacoepidemiological studies</td>
<td></td>
</tr>
<tr>
<td>Meta-analyses/systematic reviews</td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.prisma-statement.org/">http://www.prisma-statement.org/</a></td>
<td>Published in 1999</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cochrane-handbook.org/">http://www.cochrane-handbook.org/</a></td>
<td>Written by the cochrane collaboration</td>
<td>Maione et al. NDT 2011 [31]</td>
</tr>
<tr>
<td></td>
<td>Focused on systematic reviews of health research since 1993</td>
<td>Haring et al. JASN 2012 [33]</td>
</tr>
</tbody>
</table>
system-based questions related to health care organization, delivery, financing and regulation, which are important topics in CER. Finally, because observational studies can leverage information from existing datasets, they can often be completed more quickly and at lower cost than RCTs.

However, some have raised concerns about the validity of results obtained by observational studies for CER, chiefly related to issues of confounding. For example, numerous observational studies demonstrated a lower mortality risk with higher dialysis dose [37, 38]. However, the HEMO study, which randomly assigned patients on hemodialysis to higher or standard dialysis dose, found no benefit of higher dialysis dose on mortality [39]. In a clever re-analysis of the HEMO study, Greene et al. [40] examined achieved (rather than target) dialysis dose and found that the lowest quintile of achieved k/t/V was associated with a 2-fold higher risk of death compared to the middle quintile. The discrepancy between the intention-to-treat and the as-treated results of the HEMO study point toward the problem of confounding found in many observational studies. Despite this example, a systematic review found generally very good correlation among the results from randomized and non-randomized studies [41]. Moreover, advances in statistical methods in observational studies have led to continued improvement in the ability to control for confounding.

Methods to reduce confounding in observational studies

A key concern in observational studies is confounding by indication or channeling bias, particularly when examining a pharmaceutical agent [42]. Unlike in an RCT, where essentially a flip of a coin determines treatment with Drug A versus Drug B, in the observational study, a myriad of physician and patient factors may determine who receives one therapy over another. Sicker patients may be more likely to receive certain treatments than healthier patients. Even if the treatment is effective, the sicker patient will be more likely to die than the healthier patient, leading to estimates showing a detrimental effect of a treatment that prevalent users would not have been observed. How- ever, a major limitation to the new user approach is that the sample sizes are often drastically reduced with exclusion of prevalent users.

Another approach to limit confounding by indication is propensity score methods [44, 45]. The propensity score is estimated by combining a large number of variables into a model that predicts the likelihood of receipt of Drug A versus Drug B, for example. Each patient’s distinct pattern of characteristics is then combined with the coefficients of the propensity score model to calculate the propensity score. Assuming adequate model calibration and fit, participants who received Drug A will be matched to a participant with a nearly identical propensity score but who instead received Drug B. The goal is to construct a cohort with closely matched measured characteristics, with the hope that by achieving balance in the measured variables, unmeasured variables and confounders will also be balanced. This has led some to refer to propensity score matching as ‘pseudo-randomization’, although the fact remains that residual confounding can still exist after even very tight propensity score matching. Recently, there has been interest in leveraging advances in computer technology for propensity score development, a technique known as high-dimensional propensity scores [46]. This method, developed for large claims-based data, uses a multistep computer algorithm to adjust for hundreds of confounders, including interaction terms or other variables that normally might be overlooked by a researcher using a more traditional propensity score approach.

Instrumental variables, originally developed in the field of econometrics, have been applied to observational CER as well [47]. An instrumental variable is a characteristic which is associated with treatment allocation, and has no other associations with the outcome except through treatment allocation. The quintessential instrumental variable is the ‘coin flip’ that determines the treatment group at randomization in the RCT. Although finding a suitable instrument can be challenging, several recent studies in nephrology have used this approach, often using provider preference at the physician or facility level as an instrumental variable [48, 49]. For example, Bond et al. [50] examined the association of dialyzer reuse and mortality. The instrumental variable used in that analysis was the dialysis center, as certain centers had 0% reuse while other centers had >95% re-use. Assuming that patients were essentially randomly allocated to certain dialysis centers and thereby to dialyzer reuse or not, the dialysis center served as a valid instrumental variable for that analysis.

Confounders that change over time are another challenge to observational CER, particularly in nephrology, where longitudinal follow-up is required for certain outcomes of interest. Marginal structural model approaches can account for situations when time-varying variables can simultaneously be confounders as well as predictors of use of the treatment of interest [51]. However, these ‘causal’ methods also rely purely on observed data and are susceptible to constant or time-dependent confounding by unobserved confounders.

As with RCTs, different evaluative tools have been developed to assess the quality of observational studies [52]. Given the particularly challenging problems of confounding when using observational studies for CER, such quality assessments are critical prior to incorporating the results from these studies into any kind of decision making. Several assessment tools are listed in Table 2, including the Good ReseArch for Comparative Effectiveness (GRACE) principles [19, 27], which are specific to CER.
**Systematic reviews and meta-analyses**

It is common for multiple studies—whether RCTs or observational studies—to have varying or even conflicting results. Systematic reviews can therefore be useful to aggregate and summarize the available evidence, which can also be used to identify evidence gaps and potential areas for future studies. Systematic reviews should explicitly state the methods used to identify relevant studies for inclusion, which should include searches not only of published studies but also abstracts, conference proceedings or personal communications with individual investigators as necessary. Systematic reviews therefore differ from narrative reviews, where the authors generally choose different studies for inclusion at their discretion.

Sometimes, individual studies may lack sufficient sample sizes and therefore lack precision in their effect estimates. Therefore, when possible, the data from the individual trials included in a systematic review can be quantitatively pooled, analyzed and summarized as a meta-analysis. However, if the studies are too heterogeneous, then quantitative meta-analysis may not be appropriate. Importantly, the quality of the meta-analysis is only as good as the quality of the individual studies; if the available trials are all of poor quality, no amount of statistical machinations can change that fact. Selected guidelines for evaluating the quality of systematic reviews and meta-analyses are listed in Table 2.

**Why do we need CER in nephrology?**

Our patients with CKD and ESRD are among those with the highest risk for adverse outcomes, with a risk of death rivaling that of patients with metastatic cancer or advanced heart failure [53, 54]. Yet, patients with CKD and ESRD are frequently excluded from RCTs, and observational studies may not always have sufficient numbers of patients with CKD to allow for meaningful analysis in these subgroups. Between 1966 and 2002, nephrology as a whole had the fewest published RCTs of 12 medical subspecialties evaluated [55], an embarrassing statistic that has not improved substantially over the past decade. Recently, the Institute of Medicine’s Committee on CER Prioritization established a list of the 100 most pressing CER questions to be asked, but only one of these topics was directly related to nephrology (the comparative effectiveness among renal replacement therapies) [7]. Clearly, there are many other important unresolved questions in nephrology, but it appears that once again we are getting left behind the other disciplines. It is time for nephrology to move from being an opinion-based specialty to an evidence-based specialty. It is imperative to set CER priorities within nephrology in a broad and coordinated effort and to conduct focused high-quality CER to fill in the evidence gaps in areas needed to inform policy recommendations, clinical guidelines or treatment choices. The ultimate goal is to provide efficient and effective health care that improves the lives of our patients with CKD. CER pertaining to kidney disease care, be it through RCTs, observational studies or systematic reviews, ought to be a top priority to help to get us there.

**Conflict of interest statement.** None declared.

**References**
