As the number of therapeutic choices grows, there is an increasing demand for information about the comparative effectiveness (CE) of treatments for various clinical conditions and patient subgroups. This information is used to guide clinical care and training, and to support decision-making about capital expenditures for equipment. Data on CE are also used to add clinical nuance to formulary decisions about benefit coverage within drug classes and drug uses (e.g., which pharmacotherapy to use first, single or combination therapy, among others) [1], across classes of similar products (e.g., generic, preferred vs nonpreferred brands), and between different types of therapies (e.g., medical or surgical).

Randomized controlled trials (RCTs) are an important tool to evaluate CE as randomization limits confounding imbalances that interfere with our ability to distinguish between modest treatment effects and biases. However, despite their elegance and enhanced internal validity, RCTs are expensive and, since they are always prospective, can be lengthy. Hence it is unlikely that there will ever be enough time or funding for every healthcare issue to be addressed using a RCT [2,3]. Physicians, payers and patients have also begun to advocate for using observational research to shape understanding about CE, since these studies can usually be conducted at a lower cost and more quickly, and can provide useful information not otherwise available about treatment effects in large populations and in subgroups of special interest, also known as treatment heterogeneity [4,5].

Note that studies in which treatment is determined by the physician and patient are often referred to as ‘observational’, ‘nonexperimental’, ‘noninterventional’ and sometimes ‘real-world’ as treatment decisions are not assigned by a randomization protocol, but instead are determined by patient characteristics, physician preferences and factors relating to treatment access and availability [6]. The term observational is used here to avoid confusion with the ‘PICO(TS)’ terminology used by the Agency for Healthcare Research and Quality and others, which addresses study planning in terms of population of patients, ‘interventions or exposures’, comparison, outcomes and sometimes, timing and setting [101]. Cohort studies are the principle tools used for observational CE research and these can be further differentiated into those that use prospective data collection and those that rely exclusively on repurposing existing data.

Prospective and retrospective observational studies are increasingly being used to assist in developing clinical guidelines, health policy and formulary decisions. For example, the Academy of Managed Care Pharmacy recently updated its formulary submission guidelines to note the value of observational research for CE, highlighting its ability to provide information about typical care settings with a focus on what is important to healthcare decision-makers, such as patients, and health plans [102]. This focus stems in part from the US Patient Protection and Affordable Care Act, which emphasizes the importance of moving beyond average treatment effects to understanding treatment heterogeneity among patients, and also stresses the value of health-related quality of life and other outcomes that
are directly relevant to patients [103]. Outside the USA, the UK’s NICE and Australia’s Life-Saving Drug Program, among others, have long histories of using data from disease registries to support payment decisions.

Despite the growing use of observational studies to support policies and formulary access, there is a persistent concern that results from such studies are always distorted owing to confounding imbalances and other threats to internal validity; such as bias, which could either exaggerate or minimize benefits and harm. For example, sicker patients are not good surgical candidates and this type of channeling or selection bias makes it difficult to study the CE of surgical and pharmacological treatments in observational settings, and many inappropriate criteria are being used to evaluate observational CE studies (e.g., arbitrary analytic requirements that are not universally appropriate) [104]. While it is easy to dismiss most, if not all, observational CE studies as inherently flawed owing to known or unknown factors that dictate treatment, choices and affect prognosis, or because investigators are not blinded to the treatments they prescribe, the cost of ignoring all such studies is high, since information is needed now that can be used to guide access to effective treatments [7].

Are there criteria for methods and data quality that can be applied generally to observational studies to distinguish those that couple good data with strong methods to facilitate appropriate comparisons and inferences?

The Good Research for Comparative Effectiveness (GRACE) initiative was created to develop tools that could be used to boost rigor and foster recognition of good-quality observational studies on CE [105]. The intended audience is those who make decisions about treatments and determinations about access to treatments, for example, patients, payers, providers and policy-makers. The initiative was launched with seed funding from the National Pharmaceutical Council and has benefited from many volunteers who helped fine tune the principles and test checklist items by rating articles. Participants in the GRACE initiative come from professional associations, government agencies, drug and device manufacturers, and academia in North and South America, Europe, Asia and Africa.

**GRACE principles**
The GRACE principles provide guidance about the essential elements of good practice for study design, conduct and reporting for observational studies of CE [8]. Drafted with formal review and endorsement by the International Society for Pharmacoepidemiology (ISPE), the GRACE principles can be summarized in three questions, which are explained below:

- Were the study plans, including research questions, main comparisons, outcomes and so on, specified in advance of conducting the study?
- Was the study conducted and analyzed in a manner consistent with good practice, and reported in enough detail for evaluation and replication?
- How valid is the interpretation of CE for the population of interest, assuming sound methodology and appropriate follow-up?

A study plan should be prepared at the outset and should describe the target study population, diseases/conditions, comparators, outcomes and methods for study conduct. Use of clinically meaningful outcomes in preference to surrogate endpoints is emphasized (e.g., measuring fractures rather than bone density). Transparency about methods and reporting is encouraged so that studies can be replicated in other populations and settings.

The GRACE principles stress the importance of comparing study subjects who are similar in the characteristics that would cause them to receive the treatments of interest. For studies of medication effectiveness, one of the first challenges is to understand the indications and contraindications for use. This challenge also has some applicability to devices, diagnostics and medical procedures, though potential confounders may be less well understood in these contexts. The other major types of bias that could lead to systematic differences between treated groups should be considered. These include: channeling of treatments to sicker or healthier patients (also called confounding by indication/contraindication, as discussed previously); differences in how treated groups are selected, assessed and followed over time; misclassification of treatments or outcomes; and other systematic differences in care aside from the diagnostic tool or treatment under study. These various systematic differences are present to some degree in all observational research, by its very nature. The degree to which bias is present depends on the specific situation. Various statistical modeling approaches, including propensity score adjustment and instrumental variables, can be used to facilitate the comparison of people
with a similar likelihood of receiving treatment [9–12]. Sensitivity analyses can help to quantify the extent to which an effect estimate may be biased by examining how the effect estimates change depending on key assumptions used in the analyses. This approach is frequently used to understand the impact of missing data, although it has many beneficial applications [13].

The validity of the interpretation depends on understanding how determinants of treatment choice are related to the expected outcomes. A hierarchy of evidence for observational CE research was developed to classify the likelihood that effect estimates may be biased. The pinnacle of the hierarchy, representing the least biased evidence, includes studies where the determinants of treatment are not related to the determinants of outcomes; for instance, an absence of confounding. For example, comparing the risks and benefits of treatment decisions that are largely driven by reimbursement practices, such as formularies that differ by health insurance plan, would enable an observational study to achieve a good balance between comparison groups because an individual’s choice of insurance plan is not usually related to drug formulary status. However, these studies may still be biased to the extent that the populations covered by health insurance plans differ in important ways, such as socioeconomic status, geography and access to care, among others, all of which could have independent effects on outcomes. Complicating this issue even further, information about the formulary status of any particular product at various points in time is rarely accessible to researchers.

The middle category of the hierarchy is used for situations in which confounding and bias may occur; however, broadly speaking, the determinants of treatment are unlikely to be related to the outcome of interest. For example, when clinical equipoise occurs and a variety of treatments are frequently used with no good evidence for one treatment over another, or when the determinants of treatment are largely known but are not likely to be related to outcomes. Another example is when there are conflicting guidelines, such as the discrepant advice on the value of using erythropoietin to treat anemia in patients with kidney disease. In this example, an article warning about the risk of heart failure from erythropoietin use was published 1 year before the National Kidney Foundation issued guidelines recommending aggressive treatment with erythropoietin in the same types of patients [14].

The lowest level in this evidence hierarchy includes studies where confounding and bias are likely. In this category, the determinants of treatment may well be independently related to the outcome(s) of interest, directly or indirectly. Despite these apparent flaws, there may still be useful evidence about CE, especially in situations where the observed effect is much larger than what would likely have resulted purely from bias, such as a doubling or greater in the apparent benefit/harm. True differences that are relatively small are less likely to be explained completely by bias, unless clear and convincing explanations can be mounted that are specific to the situation. There may also be some benefit even to having biased CE data when no other information is available, although it may be very difficult to gauge the direction and amount of bias likely to be present.

The GRACE checklist

The GRACE checklist was created to facilitate the screening of observational studies of CE. The checklist contains a set of broadly applicable questions for screening observational CE studies to identify those that are likely to produce reasonably accurate and unbiased effect estimates. An 11-item checklist was developed through literature review and consultation with experts from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), ISPE, payers, private sector and academia. Item content addresses the comparability of subjects, the quality and applicability of data on treatments and outcomes, and statistical analyses – metrics that are similar to those employed in assessing observational study quality in systematic reviews [15–18]. Checklist items include six questions about data (treatments, primary outcomes and important covariates) and five questions about methods (design, comparators and analysis, including immortal time bias and sensitivity analyses).

The checklist is focused on whether the data collected and study methods are sufficient for the study purpose(s). The checklist, instructions for use and related publications are available on the internet [105].

A fundamental challenge was to find a ‘gold standard’ against which to test checklist items. Instead of choosing a single approach to validation, we tested the checklist using three approaches with the intent of identifying consistency. Volunteers
from many countries rated published studies of drugs, medical devices and medical procedures that were identified from systematic reviews and recent publications. The volunteer ratings were compared with overall assessments of quality from published systematic reviews or based on opinions solicited from academic and industry experts. The testing was not framed in the context of any specific question (e.g., should this treatment, at this dosage, be approved for use in a particular population and situation). The checklist items performed reasonably well with this testing; however, no scoring or specific pass/fail criteria have been identified.

Discussion
What sets GRACE apart from other tools to assess the research and evidence quality of individual observational studies of CE [19–21]

Like other tools, GRACE was derived from a thorough literature review and includes many aspects of study quality identified by experts and consensus groups. However, unlike most other tools developed for this and similar purposes, the GRACE checklist has been subjected to tests that demonstrate that its items are largely evaluable; rates without much specific training can apply the checklist items to a wide variety of study questions, and the checklist items have quantitative evidence of their value in distinguishing relatively high-quality information. By contrast, ISPOR, for example, has promulgated the importance of a prespecified hypothesis as a marker of quality [22]; however, this information was not available in any of the 88 published articles reviewed in the GRACE validation activities. Other strengths of the GRACE initiative include the broad collaboration with experts in pharmacoepidemiology and outcomes research, the relatively large-scale testing by volunteers with diverse backgrounds, and the applicability to drugs, medical devices and medical procedures.

The GRACE initiative tools need further development to address specific issues and contexts, such as those relating to types of decisions faced by pharmacy, payer and other healthcare constituencies, and to further evaluate the skills needed to rate observational CE studies effectively. The appeal of this broadly applicable tool needs to be balanced by its lack of specificity to various therapeutic areas; for example, oncology presents special challenges owing to the myriad of treatment combinations and sequences used, and the increasing attention to response differentiation by biomarkers. Finding a well-accepted standard against which to test checklist items to further refine the distinguishing aspects of quality also remains an open question, since no widely accepted gold standards exist [23,24]. It may also be beneficial to conduct separate testing for safety and effectiveness, and to develop companion documents that address the applicability of more specific quality standards, such as Good Clinical Practice, to prospective observational CE studies, or to develop new detailed standards.

In its current form, the checklist may be more useful for highlighting studies that clearly fail to meet the basic attributes of sufficient quality, than using it for confirmation of quality and appropriateness. It is unlikely that a strategy for grading studies or creating an overall quality score will be developed in the near future. Although such an approach has some appeal, it is unlikely to be useful without a thorough understanding of how items should be weighted. The applicability of studies that survive this scrutiny will largely depend on the specific decision at hand since the data and methods must be appropriate for the purpose. The existing evidence in a field also needs to be taken into consideration since, for example, well-conducted RCTs are available for the specific formulation, comparison and populations of interest, then observational studies may have little else to contribute.

Hopefully, the tools developed by the GRACE initiative will promote the appropriate use of observational CE studies through understanding the key principles of good practice, recognizing data and study methods that are sufficient for the study purpose, and remembering that study quality needs to be evaluated in the context of a specific study purpose. Those interested in more serious evaluation of observational CE studies should consult information about good practice for protocols for observational studies of CE [25], methodologic guides prepared by the European Networks of Center of Pharmacoepidemiology and Pharmacovigilance [106] and should have a good understanding of epidemiology and related disciplines [26].

Conclusion
Through examination of study methods, data quality and appropriateness of observational CE studies, the GRACE initiative shows that high-quality CE studies in which strong or moderately strong effects are detected can produce
useful information about treatment benefits, risks and heterogeneity. Hopefully, even these imperfect tools can be useful in achieving a better understanding of treatment benefits, harms and heterogeneity.

Acknowledgements

NA Dryer would like to acknowledge her collaborators, P Velentgas and A Bryant from Quintiles Outcome, and R Dubois and K Westrich from National Pharmaceutical Council, and to thank the advisors and many volunteers who rated articles and gave feedback (see [10] for a list of contributors).

Financial & competing interests disclosure

Some support for the GRACE principles and checklist was provided by the National Pharmaceutical Council. NA Dryer is employed by Quintiles Outcome, a company that conducts prospective and record-based observational research on comparative effectiveness and safety. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:
- of interest
- of considerable interest
- n
- Paper of special note has been highlighted as:
- \n
- Provides a good description of payers’ needs and is rich with examples.
- Provides a list of other scales that have been used for similar purposes.
- Provides a good summary of the challenges of developing tools to rate study quality.
25 Velentgas P, Dryer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness...
Provides detailed guidance for protocol development for observational comparative effectiveness studies.


Websites


One Hundred Eleventh Congress of the USA: at the second session. www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf


GRACE. Good Research for Comparative Effectiveness. www.graceprinciples.org


Provides a short, well-written summary of important considerations for research on comparative effectiveness and safety.