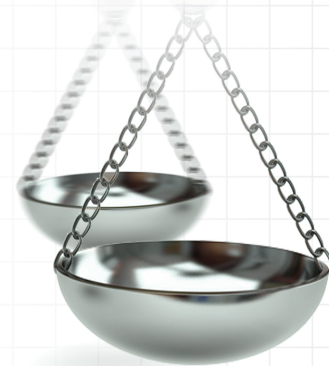




Evaluation of comparative effectiveness research: a practical tool

Debra A Schaumberg¹, Laura McDonald², Surbhi Shah¹, Michael Stokes³, Beth L Nordstrom¹ & Sreeram V Ramagopalan^{*,2}

¹Real-World Evidence, Evidera, Waltham, MA 02451, USA

²Center for Observational Research & Data Sciences, Bristol-Myers Squibb, Uxbridge, UB8 1DH, UK

³Real-World Evidence, Evidera, St-Laurent, Quebec, H4T1V6, Canada

*Author for correspondence: Tel.: +44 1895 523000; sreeram.ramagopalan@bms.com

Journal of **Comparative Effectiveness Research**

Comparative effectiveness research (CER) guidelines have been developed to direct the field toward the most rigorous study methodologies. A challenge, however, is how to ensure the best evidence is generated, and how to translate methodologically complex or nuanced CER findings into usable medical evidence. To reach that goal, it is important that both researchers and end users of CER output become knowledgeable about the elements that impact the quality and interpretability of CER. This paper distilled guidance on CER into a practical tool to assist both researchers and nonexperts with the critical review and interpretation of CER, with a focus on issues particularly relevant to CER in oncology.

First draft submitted: 11 January 2018; Accepted for publication: 6 February 2018; Published online: 21 February 2018

Keywords: cancer • comparative effectiveness research • epidemiology • immuno-oncology • methods • observational studies • review

Comparative effectiveness research (CER) compares the benefits and/or harms of health interventions in real-world settings in which care is provided to patients under routine clinical practice conditions. CER aims to improve health outcomes through the generation of evidence about which interventions are most effective for which patients under which conditions [1]. CER can include pragmatic trials, observational studies (e.g., cohort, case-control and cross-sectional), decision analytical models and systematic research syntheses [2].

The proliferation of medical record and administrative claim databases has meant that observational studies of drug treatments represent an increasing proportion of CER. Some advantages these studies have over randomized controlled trials include a larger number of patients, broader and more heterogeneous patient populations, longer follow-up periods and the ability to address questions that are not amenable to randomization. Evidence generated from these studies is increasingly used to guide clinical and policy decision making on the safety and effectiveness of healthcare interventions in the real world. However, due to the nonrandom allocation of treatments, observational studies of routine clinical care are also subject to bias that can undermine the validity of study findings and impede appropriate evidence-based decision making.

Sources of bias in CER can be minimized, however, through careful attention to study methods. National and international research organizations have developed guidelines and tools focused on observational research methods pertinent to CER. The purpose of this work was to review and distill this core guidance into a practical set of principles for researchers and other interested end users to facilitate the critical review and interpretation of CER for observational studies of drug treatments. While many guidelines for the conduct of observational research exist, there are relatively few that focus exclusively on CER. Due to the number of considerations inherent to observational CER, the authors observed a need for a concise tool for assessment of the degree to which a study follows core CER methods principles. The focus of this tool was directed predominantly toward CER in oncology, which is of increasing importance due to the rapidly advancing and complex treatment landscape in this therapeutic area. In addition, many new treatments in oncology gain accelerated approval based on early clinical trial data. In these situations, the use of CER and real-world evidence can play a particularly important role in understanding

the generalizability of trial results to patients in routine clinical practice as well as providing further insight into dosing and safety issues early in the product lifecycle [3].

Development of a concise CER methods tool

We screened national and international resources for guidelines on CER, including: Patient-Centered Outcomes Research Institute [4], International Society for Pharmacoeconomics and Outcomes Research [5], International Society for Pharmacoepidemiology [6], Agency for Healthcare Research and Quality [7], European Network of Centers for Pharmacoepidemiology and Pharmacovigilance [8], Strengthening the Reporting of Observational Studies in Epidemiology [9], Good Research for Comparative Effectiveness Initiative [10], CER collaborative [11], and Reporting of Studies Conducted using Observational Routinely collected Health Data [12]. Additionally, we conducted a brief review of electronic databases such as PubMed, MEDLINE, Google Scholar and Web of Knowledge to identify other relevant documents, including a special 2012 issue on CER in oncology published in the *Journal of Clinical Oncology* [13]. This study was not a comprehensive systematic review of all published guidelines on observational research.

We identified 14 primary sources of information (summarized in [Supplementary Table 1](#)). We extracted information on CER methods, prioritized key elements deemed to be most relevant to researchers and end users through expert consensus and synthesized the resulting content in a practical tool with the following three sections: importance of the research question; appropriateness of the data source; and rigor of the study methods. We identified specific CER study characteristics in each section and rated each along a scale of low, medium or high levels of adherence to key methodological principles. We color-coded the rating scale to provide an intuitive visual summary of study quality across these dimensions ([Table 1](#)). We then applied this tool in two case examples in oncology.

A practical tool for CER

Importance of the research question

Key consideration: the value of any CER study depends on the importance and specificity of the study hypothesis. CER should aim to address an important evidence gap (e.g., head-to-head comparisons of therapies where direct evidence of safety and effectiveness are lacking).

Are the right patients being studied?

The real-world effectiveness of a healthcare intervention can depend on many characteristics such as age, sex, race/ethnicity, disease severity, prior treatments and insurance status [14]. In oncology, these and other characteristics (e.g., tumor stage, tumor characteristics [e.g., genomics and biomarker status] and line of therapy) are likely to have an impact on outcomes. However, due to restrictive inclusion and exclusion criteria in most clinical trials, existing data on innovative therapies at the time of market approval and beyond may be limited to a small subset of the population of patients for which the treatment might be considered. This can limit generalizability (e.g., if only advanced disease is included, results may not be generalizable to patients with earlier stages of disease). CER can address gaps in evidence and inform assessments of real-world drug effectiveness, tolerability and safety. Details on inclusion and exclusion criteria and study time periods should be clearly specified and must be chosen to address the hypothesis being tested to ensure generalizability to the appropriate patient group. Additionally, to support overall conclusions and study generalizability, results can be reported for patient subgroups of interest (e.g., elderly persons, distinct ethnic backgrounds, those with poor performance status or tumor characteristics [e.g., tumor size, biomarker status and histology]).

Are the right treatments being studied?

The evolving landscape of cancer treatment has resulted in the availability of many potential treatment choices, including multidrug treatment regimens. Selection of an active comparator reflective of standard of care is an appropriate choice when evaluating real-world effectiveness of a newly approved therapy. However, standard of care may not always be obvious if there are substantial variations across clinical practice (e.g., regional or health system differences). Selection of one comparator drug or drug regimen is an option in such cases. Choosing appropriate active comparators may reduce the potential for bias, particularly confounding by indication or severity [15]. This is the tendency for the indications that can impact risk of study outcomes to also influence the prescribing tendencies of physicians (e.g., presence of high-risk disease features or overall health status). This can result in biased estimates

Table 1. A concise comparative effectiveness research methods tool.

Principles		Key considerations and degree to which principles were addressed [†]		
		High degree	Moderate degree	Low degree
1	Importance of the research question			
1.1	Are the right patients being studied?	• Characteristics of participants that may impact occurrence of the primary outcome (e.g., age, sex, tumor stage and biomarker status) are comparable to target population of interest • Important subgroups are included to support conclusions related to more targeted treatment decisions	• Most of the characteristics are comparable to the target population; any differences are not likely to impact generalizability • One or more subgroups may be absent or too small for analysis	• Characteristics of participants are different from the target population in important ways that limit generalizability of study findings
1.2	Are the right treatments being studied?	• Treatment comparisons are meaningful to address an important evidence gap • The selected treatment and active comparator are appropriate to minimize confounding by indication • Treatment parameters are appropriate and well described (timing of treatment initiation, line of therapy, dose, etc.)	• There are questions regarding the representativeness of the usual care regimens - OR • There are questions regarding potential for confounding by indication	• If compared with usual care, usual care differs compared with the target population - OR • No comparator is used increasing the likelihood of bias - OR • The comparator does not reflect a clinically meaningful choice for real-world practice - OR • There are questions regarding differential misclassification of exposure to the treatment and comparator (i.e., accuracy of treatment classification is unequal among patients with the outcome vs those without it)
1.3	Are the right outcomes being studied?	• Outcomes are clearly and consistently defined, identically measured in comparator cohorts and clinically meaningful • Inclusion of both effectiveness and safety outcomes • Primary outcomes are prespecified • Clinical outcomes are measured objectively • Validation study assessing accuracy of coding definitions is performed or cited	• Outcomes are clearly and consistently defined and identically measured; primary outcomes are prespecified but only partially address research questions - OR • Clinical outcomes are subject to judgment - OR • Validity of coding definitions for capture of outcomes not discussed	• Outcomes not clearly and consistently defined, or not measured identically - OR • Outcomes are not relevant to patients/do not address evidence gaps - OR • Outcomes do not allow assessment of benefits vs risks (e.g., no safety outcomes) - OR • Primary outcomes are not prespecified
1.4	Is the right timing being used for the study?	• Treatment parameters match the question of interest (e.g., stage of disease at treatment initiation) • Exposure and follow-up period is sufficiently long to fully assess primary outcomes	• There are questions regarding important treatment parameters, but these are mostly well matched to the research question • Follow-up time is minimally sufficient	• Timing of treatment is unknown or does not match study questions - OR • Follow-up period is not sufficient to address primary outcome
2	Appropriateness of the data source			
2.1	Does the data source contain sufficient information on treatment and outcome parameters?	• Data source contains sufficient detail to define: ◦ The appropriate indication ◦ Treatment parameters (timing of treatment start, line of therapy, dose, etc.) ◦ Confounding variables ◦ Effect modifiers (e.g., magnitude of treatment effect is different for certain individuals) ◦ Outcome measures (e.g., sufficient diagnostic and procedure codes)	• The needs of the study question are not completely met by the primary data source but steps were taken to address the limitations with supplemental information (e.g., a validation study was conducted on subset of patients to estimate the sensitivity and specificity of algorithm used to identify outcomes) - OR • Indication cannot be established definitively, but drug is approved for only one indication and off-label use is uncommon	• The data source does not meet the needs of the primary study question (e.g., there is insufficient data to identify important treatment parameters, confounding factors, etc.)
2.2	Does the study include enough patients to ensure statistical power to address a clinically meaningful effect size?	• The study provides clear details on sample size/statistical power (study power $\geq 80\%$) • The resulting number of patients and outcomes in each comparator cohort is sufficient to detect clinically meaningful differences in the primary outcomes	• Sample size calculations not provided, but size of study cohorts is sufficient to address the primary outcomes	• Regardless of reporting on sample size details, the size of the cohorts is insufficient to address primary outcomes (e.g., wide confidence intervals reducing interpretability and utility of findings)

[†]Multiple bullet points in any section not separated by 'OR' should be interpreted as 'AND'; thus all items should be met to qualify at that level.

Table 1. A concise comparative effectiveness research methods tool (cont.).

Principles	Key considerations and degree to which principles were addressed [†]		
	High degree	Moderate degree	Low degree
3. Rigor of the study methods			
3.1 Does the study methodology target new initiators of the treatment?	• New initiators are identified • Sufficient information is provided to establish: ◦ Line of therapy ◦ Washout period ◦ Timing of treatment initiation	• Prevalent users are studied but justified by rareness of outcome or use of long-term outcomes limiting number of exposed patients • No mixing of prevalent and new users • No mixing of lines of therapy	• Prevalent cohort is used when a new user cohort may have been available OR • Study mixes new and prevalent users OR • Study does not appropriately consider therapy
3.2 Are the comparator cohort(s) included in the study from the same time period as the main intervention?	• Comparator cohort selected from same time period and same data source, window of exposure is clearly defined ◦ Induction period (if used) is defined	• Historical comparison cohort is used but justified OR • Comparison cohort is selected from a different data source, but justified OR • Appropriate analytic techniques are applied to reduce confounding	• Cohort selected from different time period without providing proper rationale
3.3 Does the analysis include careful consideration and application of appropriate techniques to control for potential bias?	• Appropriate methods are used to adjust for baseline differences in confounders such as multivariable regression or propensity score methods • Person-time is correctly classified to avoid immortal time bias	• Multivariable analyses are performed to control for confounding, but there are concerns regarding model fit, number of covariates in relation to the outcome or variables included in the model OR • There are concerns regarding important unmeasured confounders AND • Person-time is correctly classified to avoid immortal time bias	• Insufficient control for confounding (e.g., no adjustment method used) OR • Important unmeasured confounders OR • Person-time is incorrectly classified resulting in immortal time bias
3.4 Are sensitivity analyses performed to assess robustness of the study findings?	• Multiple sensitivity analyses were performed to examine impact of alternate assumptions	• Some sensitivity analyses are performed	• Very limited or no sensitivity analyses are performed

[†]Multiple bullet points in any section not separated by 'OR' should be interpreted as 'AND', thus all items should be met to qualify at that level.

of intervention-associated outcomes simply because of a differentially higher or lower baseline risk of the outcome among patients who receive a particular therapy. For instance, oncologists may selectively prescribe a single-agent treatment regimen to those who are in overall poor health instead of a combination (e.g., >1 drug) regimen. In this example, the effectiveness of the combination regimen may be artificially inflated relative to the single-agent regimen. In other cases, bias may occur where an active treatment is compared with palliative care, or in instances where tolerability differences, adverse event profiles or other factors differ among treatment choices. Under some circumstances, however, it may be challenging to select an active comparator, or the evidence gap may involve active therapy versus observation. For example, in the case of slow-growing tumors such as prostate cancer, observation is one plausible treatment strategy that involves waiting and closely monitoring the progression of the disease. In such cases, comparing this strategy with active treatments could be appropriate, with careful attention to control for confounding. However, if the data source does not contain enough information on important confounding factors, the waiting/monitoring strategy may appear more effective simply because physicians are reserving this strategy for those with a better prognosis.

Are the right outcomes being studied?

Study outcomes must be clearly and consistently defined, clinically meaningful and identically measured. Primary outcomes should be prespecified and defined in detail. Objective measures of clinical outcomes are preferred over outcomes subject to clinical judgment. The accuracy and comprehensiveness of diagnostic and procedure codes used to capture clinically meaningful outcomes are important measures of the quality of a CER study. These should be applied in the same manner in both the treatment and comparison cohorts to avoid bias. In prostate cancer, for example, identification of treatment-related urinary incontinence may involve both a set of diagnostic and procedure codes, such that the exclusion of any relevant code may impact estimates of treatment-associated harms. Validation studies or assessments examining the ability of the coding definition to accurately detect an outcome should be considered whenever feasible, or published validation work should be referenced.

In addition to the accuracy of outcome ascertainment, the clinical meaningfulness of CER outcomes (i.e., how well the outcome provides the evidence needed for patients and their doctors to make treatment-related decisions) is paramount. In oncology, survival (progression-free, cancer-specific or overall) is commonly studied. Other outcomes such as adverse events, side effects and patient-reported outcomes could be of importance to patients. Economic measures including healthcare resource utilization and costs are also often of interest to payers and health system administrators.

Is the right timing being used for the study?

The timing of the study treatments in relation to factors such as stage of disease and tumor size should be specified and relevant to the research questions. Certain therapies may have delayed effects that could alter assessments of long-term benefit versus risk. The follow-up period should be consistent with the investigators' understanding of proposed biological mechanisms that will impact the time to outcome events (e.g., if the gap is related to lack of long-term outcomes, then the period should be long enough to capture those).

Appropriateness of the data source

Key consideration: a crucial factor for any CER protocol is the capability of a data source to enable the study of enough patients with sufficient detail on treatment, disease and outcome measures, information on important potential confounding factors, as well as a sufficient length of follow-up time.

Does the data source contain sufficient information on treatment & outcome parameters?

CER often relies on analysis of data that were not specifically collected for research, such as data obtained as part of routine medical care or administrative claims. These data can increase the likelihood of bias, particularly where certain variables (representing important confounders or effect modifiers) are not available, and where treatments or outcomes are measured with error [16]. For example, some data sources may not routinely capture important features such as line of therapy, cancer stage, biomarkers (e.g., genetic tests) and this may impact the quality and clinical interpretability of CER. Consequently, it is important to understand the strengths and limitations of data sources, and to consider strategies to address them, such as replication of a study using more than one data source or supplementation of available data with information from external data sources. For example, linkages between claims databases, electronic health records and external cancer registries can provide researchers with information that can be used to validate assumptions regarding tumor stage, line of therapy and other important variables. In some cases, it may be possible to augment the information in a claims database through the collection of information from patients' medical charts. In head and neck cancer, for instance, infection with human papilloma virus is an important prognostic indicator that is usually not included in claims data but is routinely documented in patients' medical charts.

Does the study include enough patients to ensure statistical power to address a clinically meaningful effect size?

A larger sample size increases precision in estimates, provides more statistical power to detect smaller magnitude differences that can be important at the population level and increases the confidence in interpretability and applicability of study results to clinical decision making [14]. A study should provide a clear description of sample size/power estimation, and reviewers should determine whether the underlying assumptions are reasonable (e.g., estimated rates of the outcomes are derived from peer-reviewed literature and can be reasonably assumed to apply to the study population; study power is at least 80%; and the magnitude of the effect of the intervention is of clinical relevance). In cases from the literature where no detail is provided on sample size or study power estimation, the width of the confidence intervals generated around the effect estimates will provide a sense of the range of values in which the true effect is likely to exist.

Rigor of the study methods

Key consideration: the gold standard for observational CER studies employs a study design that: selects new users of therapy and a similar group of new users of an active comparator therapy from the same period; applies appropriate methods such as multivariable modeling, propensity score approaches or instrumental variable analysis to control for confounding; and conducts sensitivity analyses to understand the robustness of study findings to variations in underlying assumptions.

Does the study methodology target new initiators of the treatment?

Among the most important threats to study validity in observational CER are three types of bias that can result from the selection of prevalent users of treatment. These types of bias include: an underestimation of event rates if events occurring when therapy is first initiated cannot be ascertained (e.g., ‘depletion of susceptibles’ bias, where a cohort is depleted of all susceptible individuals experiencing an early event); confounding by risk factors that are modified by the study drugs (which occurs when a risk factor that is affected by the treatment is included as a ‘baseline’ risk factor despite being measured after treatment initiation); and the so-called healthy adherer bias which involves the selection of people who tolerated the treatment and, most likely, for whom treatment appeared to be effective [17]. Restriction of CER to new initiators of therapy (or starters of a new course of therapy) is the most important design feature to limit these sources of bias. The new user design emulates the interventional part of a randomized clinical trial where the treatment arms are restricted to patients who are newly prescribed one of the study drugs. Essentially, patients are followed from initiation of a therapy until the occurrence of specified study outcomes or end of the study period. In addition, there is usually a minimum period of nonuse (washout) prior to the initiation of therapy [18]. In cases where the new user approach may not be suitable (such as in questions pertaining to long-term users), careful consideration of potential biases introduced by prevalent users, as outlined above, are needed.

Is the comparator cohort(s) included in the study from the same time period as the main intervention?

Use of a historic control may lead to confounding by factors that vary over time [19]. Also known as chronology bias, this situation can arise when long-term trends within the healthcare system affect the severity of the target condition, induce changes in outcome rates or alter how clinicians diagnose disease or identify outcomes. Consequently, selection of a concurrent comparison group is preferred. However historic controls may sometimes be the only option (e.g., when there is rapid high uptake of a new treatment [19]) such as may occur with the approval of a breakthrough therapy. If historic controls are used, they should be as similar to the treatment cohort as possible.

Does the analysis include careful consideration & application of appropriate techniques to control for potential bias?

In nonrandomized studies, baseline differences between the treatment and comparator cohorts could result in confounding [20]. Using observational data for CER therefore requires design features (described above) and analytic strategies to control for confounding, such as traditional multivariable regression techniques, propensity score techniques [21] or instrumental variable analysis [22,23]. When the number of outcome events per covariate included in the regression model is sufficient (e.g., ten or more), traditional multiple regression is generally reasonable [24]. In cases of many covariates, it may be tempting to use an automated variable selection method such as stepwise regression. However, such automated procedures apply subjective rules and generally do not align well with the considerations of confounding. Other analytical strategies such as propensity score methods can adjust for large numbers of covariates in CER, and often perform better than logistic regression when the outcome is relatively rare (e.g., seven or fewer events per confounder) [25]. Additional benefits of propensity score methods include their focus on understanding indications for drug selection, the ability to test for interactions between propensity of treatment and drug effects on outcomes and correction for some unmeasured confounding with propensity score calibration. Alternatives such as disease risk scores may be a good choice for analysis of a common outcome and rare or multiple exposures, and instrumental variable analysis is one of several possible approaches to address unmeasured confounders [23]. The forms of the study outcomes, treatments and covariates determine the specific statistical methods to be used. For example, CER studies often investigate time-to-event data with variable follow-up and censoring of outcomes; these are commonly analyzed using Cox proportional hazards regression.

Many observational studies use a cohort approach to emulate randomized controlled trial designs. However, if the exposure period is not correctly classified, there is a risk of immortal time bias [26]. Consider a study in which cohort entry is based on a clinical event such as a diagnosis of lung cancer, with patients followed from the date of diagnosis until death. If the study aims to compare chemotherapy use versus no chemotherapy, and defines exposure by administration of a chemotherapy agent within 90 days of the initial diagnosis, then the time between cohort entry (date of lung cancer diagnosis) and administration of the first chemotherapy agent is immortal since patients must survive this period to receive their first prescription for chemotherapy. Since this immortal period has been misclassified, the chemotherapy-use group will have an artificial survival advantage over unexposed patients [26]. A way to avoid immortal time bias in this example is to align the start of study follow-up with the initiation of chemotherapy, rather than the initial lung cancer diagnosis or more generally to start follow-up only when all

Table 2. Case example #1 illustrating application of the comparative effectiveness research principles[†].

	CER principles	Quality indicator	Discussion
1	Study design and cohort selection		
1.1	Are the right patients being studied?	‡	South Korea provides universal health insurance coverage and the selected database in this study covers 97% of the Korean population, so this is representative of the patient population of interest. Pancreatic cancer is of importance as it ranks as the fifth-highest cause of cancer-related mortality in Korea and 5-year OS has not changed in the past decade
1.2	Are the right treatments being studied?	§	The comparator GEM was the previous standard approved treatment option in Korea for advanced pancreatic cancer, and therefore is a relevant comparator. Addition of other drugs to GEM had previously not shown any added survival benefit, until a Phase III trial demonstrated a modest increase in 1-year survival from 17 to 23% GEM-E [30], leading to market authorization and reimbursement for GEM-E. However, with this modest level of survival advantage, there were questions regarding the clinical meaningfulness of a 2-week improvement in survival, as well as regarding a possible effectiveness gap once GEM-E was used in a wider group of patients. We downgraded the study, however, as it failed to include information on treatment parameters. For example, information on dose, number of cycles and treatment length are not described, although this information is either available in the data source or can be inferred. In addition, the authors cite other available therapies in the discussion section as being commonly used, and for which larger survival advantages have been demonstrated vs GEM, yet these therapies were not included in the study
1.3	Are the right outcomes being studied?	§	The primary outcomes were OS and medical costs per patient, which were clearly defined, meaningful and identically measured among treatment cohorts. However, additional clinical outcomes were not included, and no patient-centered outcomes or safety outcomes were studied. Inclusion of additional outcome measures could have provided further information to inform the overall benefits vs risks of the two treatment regimens
1.4	Is the right timing being used for the study?	‡	The period of the study was defined to align with market approval for GEM-E. Patients were followed for 3 years or until the date of death or the end of the study period (31 December 2013), whichever came first. This follow-up period seems to be appropriate considering the 5-year survival rates in advanced pancreatic cancers is 1–3% [31]
2	Data source		
2.1	Does the data source meet the needs of the study aims?	‡	The NHIS database meets the needs of the study aims for several reasons: due to universal health insurance coverage in South Korea, the population is representative (97%); NHIS includes relevant information on various demographic and clinical characteristics, procedures and prescriptions, payment information, etc. which was needed to answer the research question; effectiveness of new treatments in real-world practice can be studied well with this data source
2.2	Does the study include a sufficient number of patients to ensure statistical power to address a clinically meaningful effect size?	§	The final analysis includes 4267 cases of pancreatic cancer treated with GEM or GEM-E, which is higher than the ~1000 patients required to detect differences in OS between groups with ≥80% power for hazard ratios ≤0.80 and 70% probability of an event. However, the study did not provide information on a power calculation, nor a prespecified magnitude for the treatment difference the study would be powered to detect. Further, no confidence intervals were provided
3	Rigor of the study methods		
3.1	Does the study methodology target new initiators of the treatment?	§	The study included new initiators in the GEM and GEM-E treatment cohorts. Patients with pancreatic cancer were included only if they began first-line therapy with GEM-E or GEM between 1 January 2007 and 31 December 2012. Patients with a history of receiving GEM before 2007 were excluded, while GEM-E has been reimbursed since 2006. Further, patients with prior radiotherapy or surgical treatment were also excluded. Although new initiators were selected, the algorithm used for identification of first-line treatment with either GEM or GEM-E for pancreatic cancer is not described. Furthermore, details regarding validation and accuracy of the algorithm are not provided
3.2	Are the comparator cohort(s) included in the study from the same time period as the main intervention?	‡	Patients in both the GEM and GEM-E cohorts were selected from the same period between 2007 and 2012
3.3	Does the analysis include careful consideration and application of appropriate techniques to control for potential bias?	¶	Differences in baseline characteristics were assessed, and p-values were shown. The authors used multivariate Cox proportional hazards models to adjust for differences in sex, age and comorbidity (using the Charlson Comorbidity Index) for the outcome of overall mortality. However, there could be other tumor-related factors, such as tumor stage, tumor grade or tumor size which were not included, and time to treatment since diagnosis could have also been considered. Further, the analysis of cost data was not adjusted for baseline differences between the treatment cohorts. Person time should be correctly allocated to avoid immortal time bias since erlotinib was required to have been given simultaneously with GEM at the index date, and follow-up appears to start on date that the index GEM treatment was administered
[†] Shin <i>et al.</i> , 2016 was a retrospective observational study comparing GEM-E vs GEM alone among South Korean patients with pancreatic cancer. Quality indicator: green: high degree; yellow: moderate degree; red: low degree. ‡ Green. § Yellow. ¶ Red. CER: Comparative effectiveness research; GEM: Gemcitabine; GEM-E: Gemcitabine and erlotinib; NHIS: National Health Insurance Service (South Korea); OS: Overall survival. Data taken from [29].			

Table 2. Case example #1 illustrating application of the comparative effectiveness research principles[†] (cont.).

CER principles	Quality indicator	Discussion
3.4 Are sensitivity analyses performed to assess robustness of the findings?	¶	One sensitivity analysis was performed by extending the follow-up duration up to 5 years for OS. However, in such a rich data source, with potential to link to cancer registry data [32], there could be additional sensitivity analyses that could have been undertaken. For example, one of the limitations discussed in the study was the inclusion of patients only if they received a histological or cytological diagnosis within 1 year before the index date. While this is likely to increase specificity of the indication, it could have reduced sensitivity. Consideration of cases that did not meet this inclusion criterion could have been conducted. Analysis stratifying on available prognostic factors could have been considered. At last, in relation to subgroups, there were significant differences at baseline in, for example, the sex distribution within the GEM vs GEM-E groups with more men receiving GEM-E. Further analyses to explore such differences may have led to other insights that could be reveal new hypotheses

[†]Shin *et al.*, 2016 was a retrospective observational study comparing GEM-E vs GEM alone among South Korean patients with pancreatic cancer.

Quality indicator: green: high degree; yellow: moderate degree; red: low degree.

¶Green.

§Yellow.

¶Red.

CER: Comparative effectiveness research; GEM: Gemcitabine; GEM-E: Gemcitabine and erlotinib; NHIS: National Health Insurance Service (South Korea); OS: Overall survival.

Data taken from [29].

study inclusion criteria have been met. An alternative solution is to use time-varying analyses, where all patients are considered unexposed from the time of initial lung cancer diagnosis until the start of chemotherapy. Due to the risk of bias in observational CER studies and the limitations inherent in available methods to mitigate it (e.g., inability to fully control for unmeasured confounding), results should always be interpreted with caution.

Were sensitivity analyses performed to assess robustness of the findings?

All study findings are the result of underlying assumptions made in the design, conduct and analysis of the study. Accordingly, the validity of the inferences that can be drawn will depend on the extent to which these assumptions are valid [27]. Performance of sensitivity analyses that consider alternative assumptions regarding: the data source; criteria to define study cohorts; unmeasured confounders; exposure definitions; outcome definitions; analytical assumptions; and statistical methods can be used to investigate the robustness of study findings to various assumptions and provide a picture of the consistency of an observed result in terms of direction and magnitude. Thus, sensitivity analyses may involve replication of the study in an alternative data source, use of one or more alternative algorithms for cohort selection, treatment and outcome definitions and consideration of alternative choices for analysis. In fact, sensitivity analysis has proven to be an important and consistent predictor of study quality [28].

Additionally, if the biologic response or exposure is expected to differ in certain subgroups (e.g., based on age, race/ethnicity, genetics, baseline severity and prior treatment), consideration should be given to whether these subgroups should be analyzed separately. The research should identify segments of the study population for whom there are concerns about generalizability from prior research. Although many CER studies will lack power to examine all potential subpopulations of interest, if sufficient data are available to define such subpopulations, it can be useful to perform exploratory analyses to examine whether the best choice of therapy may differ for certain subgroups.

Case example #1: erlotinib plus gemcitabine versus gemcitabine for pancreatic cancer: real-world analysis of Korean National Database

A retrospective observational study conducted by Shin *et al.* in 2016 [29] compared erlotinib plus gemcitabine (GEM-E) versus gemcitabine (GEM) alone among South Korean patients with pancreatic cancer. The study was selected relatively arbitrarily from a list of CER studies in cancer examining effectiveness of a tyrosine kinase inhibitor. We evaluated this study by applying the set of principles discussed in this targeted review and outlined those findings in Table 2.

Table 3. Case example #2 illustrating application of the comparative effectiveness research principles[†].

	CER principles	Quality indicator	Discussion
1	Study design and cohort selection		
1.1	Are the right patients being studied?	‡	The SEER-Medicare data are a large population-based source covering ~26% of the US population. Although only Medicare-eligible patients were included, the majority of people with gastric cancer (60%) are ≥65 years old at diagnosis, so the patient population of interest is well represented. Use of the SEER cancer registry also provides detailed clinical information related to histology and stage that is necessary for identification of the stage IB-III gastric adenocarcinoma target population. Important subgroups are included
1.2	Are the right treatments being studied?	§	Use of gastrectomy alone as the relevant comparator is appropriate as the current standard of care is addition of perioperative chemotherapy to gastrectomy. Although recent RCTs have shown an added survival benefit in favor of adjuvant chemotherapy following gastrectomy [34,35], utilization of adjuvant chemotherapy remains low and has not been studied extensively in real-world settings following guideline adoption in the general population The study was downgraded as it did not contain information on treatment parameters such as length of treatment, dose, number of chemotherapy cycles and use of specific adjuvant chemotherapy regimens. In addition, the authors state that the only oral agent included was capecitabine since Medicare Part D data were not available for the full follow-up period. Information on the total dose of radiation, which is included in the authors' definition of adjuvant chemotherapy is not available in the SEER-Medicare database
1.3	Are the right outcomes being studied?	§	The primary outcome was all-cause death, which was clearly defined and identically measured across treatment cohorts. Cause-specific death was also cited as a secondary outcome in the methods. However, no results for cause-specific death appear to be reported. Validation studies have also reported mixed findings on the reliability of the cause of death from death certificates, raising concerns about the utility of this measure in cancer research [36–39] In addition, no outcomes related to treatment safety or tolerability were studied, which would have provided additional information regarding treatment risks
1.4	Is the right timing being used for the study?	‡	The study enrolled patients diagnosed with gastric cancer between 2002 and 2009, which was appropriate considering the study aimed to evaluate patients receiving adjuvant chemotherapy following its addition to the NCCN guidelines in 2002. All patients were followed for a minimum of 1 year or until the date of death (31 December 2010). In addition, information on the date of the initial gastric cancer diagnosis was available in SEER, allowing for correct ascertainment of individuals receiving gastrectomy within 6 months of diagnosis
2	Data source		
2.1	Does the data source meet the needs of the study aims?	‡	The database meets the needs of the study objectives for these reasons: SEER-Medicare is a large, population-based source of data on fee-for-service Medicare enrollees; important information on baseline demographic characteristics and procedures were sufficient to answer the research questions; and availability of important clinical information in SEER such as date of diagnosis, death, stage and tumor location for use in multivariate models to account for biases related to treatment selection
2.2	Does the study include a sufficient number of patients to ensure statistical power to address a clinically meaningful effect size?	§	The analysis included 1519 cases of gastric cancer patients receiving either gastrectomy or gastrectomy plus adjuvant chemotherapy, which is higher than the ~1300 patients required to detect differences in OS between groups with ≥80% power for hazard ratios ≤0.80 and 50% probability of an event. However, the study did not include a power calculation, nor a prespecified magnitude for the treatment difference the study would be powered to detect. Confidence intervals were reported
3	Rigor of the study methods		
3.1	Does the study methodology target new initiators of the treatment?	‡	New initiators were included in both the treatment cohorts. The algorithm for classifying patients into treatment groups was clearly explained and facilitated by having a clinically confirmed date of initial gastric cancer diagnosis in SEER. Identification and exclusion of patients receiving neoadjuvant chemotherapy (<i>prior</i> to surgery) was implemented to ensure only patients receiving adjuvant chemotherapy (within 90 days <i>following</i> gastric surgery) were included
3.2	Are the comparator cohort(s) included in the study from the same time period as the main intervention?	‡	Patients in both treatment cohorts were selected from the same time period between 2002 and 2009. Additionally, baseline analyses of year of diagnosis were also performed and did not indicate a strong physician preference to prescribe one treatment type over the other during specific years of the study
Quality indicator: green: high degree; yellow: moderate degree; red: low degree.			
[†] Yeh <i>et al.</i> , 2017 was a comparative study of gastrectomy only vs gastrectomy plus adjuvant chemotherapy in elderly patients in the USA with stage IB-III gastric adenocarcinoma.			
‡ Green.			
§ Yellow.			
¶ Red.			
CER: Comparative effectiveness research; NCCN: National Comprehensive Cancer Network; OS: Overall survival; PFS: Progression-free survival; RCT: Randomized controlled trial; SEER: Surveillance, Epidemiology and End Results.			
Data taken from [29,33].			

Table 3. Case example #2 illustrating application of the comparative effectiveness research principles[†] (cont.).

CER principles	Quality indicator	Discussion
3.3 Does the analysis include careful consideration and application of appropriate techniques to control for potential bias?	§	Crude differences in baseline characteristics were assessed. Several Cox proportional hazards models were used to adjust for differences in baseline and clinical characteristics. The first was a univariate model including treatment group only. Subsequent models included the addition of all baseline covariates as well as models implementing various propensity score analyses. Immortal time bias was minimized by excluding patients who died within 90 days of a gastrectomy, the authors also note that survival was calculated starting 90 days after gastrectomy. The models were adjusted for appropriate baseline characteristics including clinical information such as stage, Charlson comorbidity, tumor location, Lauren classification and number of lymph nodes resected The study was downgraded because information on the specific adjuvant chemotherapy agents received, as well as number of cycles, and duration of adjuvant chemotherapy was not presented. Additionally, no information on subsequent therapies administered following initial treatment is provided. This information could have been helpful to put the main results into more context. For example, was the adjuvant chemotherapy group receiving guideline concordant care? Was there any major variation in subsequent lines of therapy that could have explained further differences in survival between groups beyond the initial treatments provided?
3.4 Are sensitivity analyses performed to assess robustness of the findings?	‡	Several sensitivity analyses were performed including subgroup analyses by cancer diagnosis, age and tumor location. The impact of defining adjuvant therapy on the basis of 2 or 4 months (vs 3 months in the base case) from the date of gastrectomy and alternative definitions of adjuvant therapy that included any adjuvant chemotherapy (regardless of receipt of radiation therapy), only adjuvant chemotherapy and only adjuvant radiation therapy was also examined. These additional analyses enabled the authors to confirm that varying the adjuvant chemotherapy treatment window from 2 to 4 months had no effect on results, and similarly including radiotherapy or not including it in the adjuvant treatment definition also had no impact. At last, the authors were also able to explore the effect of radiotherapy alone (without chemotherapy) as adjuvant therapy

Quality indicator: green: high degree; yellow: moderate degree; red: low degree.

[†]Yeh *et al.*, 2017 was a comparative study of gastrectomy only vs gastrectomy plus adjuvant chemotherapy in elderly patients in the USA with stage IB-III gastric adenocarcinoma.

[‡]Green.

[§]Yellow.

[¶]Red.

CER: Comparative effectiveness research; NCCN: National Comprehensive Cancer Network; OS: Overall survival; PFS: Progression-free survival; RCT: Randomized controlled trial; SEER: Surveillance, Epidemiology and End Results.

Data taken from [29,33].

Case example #2: comparative effectiveness of adjuvant chemoradiotherapy after gastrectomy among older patients with gastric adenocarcinoma: a SEER-Medicare study

A retrospective cohort study by Yeh *et al.* in 2017 compared gastrectomy only versus gastrectomy plus adjuvant chemotherapy in elderly patients with stage IB-III gastric adenocarcinoma in the USA (Table 3) [29,33]. This study was evaluated similarly to case example #1, using the principles outlined in this review.

Conclusion

The elements that impact the quality and interpretability of CER provide valuable knowledge for researchers and end users of CER output when determining how to best use medical evidence. The development of a tool for reviewing CER methodologies/findings, which highlights core principles for consideration, addresses a key challenge faced by these groups. This new set of guidelines, with a particular focus on oncology, enables both researchers and nonexperts to use CER evidence with more confidence, and as a result, to be more effective in making decisions regarding treatments.

Future perspective

This paper develops a set of key methodological considerations to enable the critical review and interpretation of CER studies, with a focus on observational studies of drug treatments. These types of observational CER studies are subject to sources of bias that can be minimized through careful application of pharmacoepidemiologic principles for design and analysis. Regulatory bodies are defining guidance for use of real-world evidence in regulatory decision making, and real-world, observational CER studies are likely to become more prominent in guiding clinical and policy decision making in oncology care. It is therefore crucial that all stakeholders are familiar with methodological principles to appropriately assess the strengths and limitations of CER evidence. Improvements in health information technology will also lead to the collection and integration of more detailed clinical information into real-world data sources. This, alongside further enhancements to CER methodology and agreement on appropriate research methods, will ultimately enable more informed decisions on the use and adoption of new health technologies.

Executive summary

- Comparative effectiveness research (CER) compares the benefits and/or harms of interventions aimed at the prevention, diagnosis, treatment and/or monitoring of health conditions in real-world settings.
- This targeted review was undertaken to create a practical set of principles to support the critical review and interpretation of CER protocols and literature.
- We extracted information on CER methods from 14 primary sources and used expert consensus to prioritize the most relevant elements. The resulting content was synthesized in a CER evaluation tool with the following three sections: importance of the research question; appropriateness of the data source; and rigor of the study methods.
- Key considerations for CER methods include:
 - **Importance of the research question:** the ultimate value of any CER study depends on the importance and specificity of the study hypothesis. CER should aim to address an important evidence gap (e.g., head-to-head comparisons of therapies where direct evidence of safety and effectiveness are lacking).
 - **Appropriateness of the data source:** a crucial factor for any CER protocol is the capability of a data source to provide sufficient detail on treatment, disease and outcome measures, important potential confounding factors, as well as a sufficient length of follow-up time to enable the study of enough patients.
 - **Rigor of the study methods:** the gold standard for observational CER studies employs a study design that selects: new users of therapy and a similar group of new users of an active comparator therapy from the same period; the application of appropriate methods such as multivariable modeling, propensity score techniques and instrumental variable analysis to control for confounding; and the conduct of sensitivity analyses to understand the robustness of study findings to variations in underlying assumptions.
- The purpose of CER is to improve the health of patients through understanding the impact of treatment strategies in a real-world setting. In this paper, the core principles of CER were summarized and a tool was developed to highlight important CER study considerations.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/cer-2018-0007

Acknowledgements

The authors thank S Li and M Ulvestad for their critical review during the preparation of this manuscript.

Financial & competing interests disclosure

This work was supported by funding from Bristol-Myers Squibb. The authors have 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.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: ●● of considerable interest

1. Federal Coordinating Council for Comparative Effectiveness Research (FCCCER). *FCCCER definition of Comparative Effectiveness*. US National Library of Medicine (2017). <https://osp.od.nih.gov/wp-content/uploads/FCCCER-report-to-the-president-and-congress-2009.pdf>
2. Price-Haywood EG. Clinical comparative effectiveness research through the lens of healthcare decision makers. *Ochsner J.* 15(2), 154–161 (2015).
3. Blumenthal GM, Kluetz PG, Schneider J, Goldberg KB, McKee AE, Pazdur R. Oncology drug approvals: evaluating end points and evidence in an era of breakthrough therapies. *Oncologist* 22(7), 762–767 (2017).
4. Patient-Centered Outcomes Research Institute (PCORI). Patient-Centered Outcomes Research Institute (PCORI) (2017). www.pcori.org/
5. International Society for Pharmacoeconomics and Outcomes Research (ISPOR). The leading global scientific and educational organization for HEOR and its use in decision making to improve health (2017). www.ispor.org/

6. International Society for Pharmacoepidemiology. International Society for Pharmacoepidemiology (ISPE), Home Page (2017). www.pharmacoepi.org
7. Agency for Healthcare Research and Quality (AHRQ). Agency for Healthcare Research and Quality (AHRQ). Home Page (2017). www.ahrq.gov/
8. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Home Page (2017). www.encepp.eu/
9. STROBE. STROBE Statement. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) (2017). www.strobe-statement.org/index.php?id=strobe-home
10. GRACE. Good Research for Comparative Effectiveness Initiative (GRACE) (2017). www.graceprinciples.org/
11. National Pharmaceutical Council (NPC). CER Collaborative Initiative (2017). www.npcnow.org/issue/cer-collaborative-initiative
12. Benchimol EI, Smeeth L, Guttman A *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* 12(10), e1001885 (2015).
13. Browman GP. Special series on comparative effectiveness research: challenges to real-world solutions to quality improvement in personalized medicine. *J. Clin. Oncol.* 30(34), 4188–4191 (2012).
14. Dorey FJ. Statistics in brief: statistical power: what is it and when should it be used? *Clin. Orthop. Relat. Res.* 469(2), 619–620 (2011).
15. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care* 48(6 Suppl.), S114–S120 (2010).
- **Reviews approaches to control for confounding in pharmacoepidemiology database research.**
16. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects – advantages and disadvantages. *Nat. Clin. Pract. Rheumatol.* 3(12), 725–732 (2007).
17. Lund JL, Richardson DB, Sturmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr. Epidemiol. Rep.* 2(4), 221–228 (2015).
- **Reviews methodology to implement the active comparator, new user study design.**
18. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am. J. Epidemiol.* 158(9), 915–920 (2003).
19. Setoguchi S, Gerhard T. Comparator Selection. Chapter 5. In: *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Velentgas P, Dreyer NA, Nourjah P *et al.* (Eds). Agency for Healthcare Research and Quality. Rockville, MD, USA (2013).
20. Groenwold RH, Hoes AW, Hak E. Confounding in publications of observational intervention studies. *Eur. J. Epidemiol.* 22(7), 413–415 (2007).
21. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin. Pharmacol. Toxicol.* 98(3), 253–259 (2006).
22. Tan FES. Confounding in (non-) randomized comparison studies. *OA Epidemiol.* 1(3), 21 (2013).
23. Streeter AJ, Lin NX, Crathorne L *et al.* Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J. Clin. Epidemiol.* 87, 23–34 (2017).
24. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 49(12), 1373–1379 (1996).
25. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol.* 158(3), 280–287 (2003).
26. Suissa S. Immortal time bias in pharmaco-epidemiology. *Am. J. Epidemiol.* 167(4), 492–499 (2008).
27. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. AHRQ. Agency for Healthcare Research and Quality. Rockville, MD, USA (2013).
- **Contains detailed information on best practices for comparative effectiveness research (CER) protocol development.**
28. Dreyer NA, Bryant A, Velentgas P. The GRACE checklist: a validated assessment tool for high quality observational studies of comparative effectiveness. *J. Manag. Care Spec. Pharm.* 22(10), 1107–1113 (2016).
29. Shin S, Park CM, Kwon H, Lee KH. Erlotinib plus gemcitabine versus gemcitabine for pancreatic cancer: real-world analysis of Korean national database. *BMC Cancer* 16, 443 (2016).
30. Moore MJ, Goldstein D, Hamm J *et al.* Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a Phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J. Clin. Oncol.* 25(15), 1960–1966 (2007).
31. American Cancer Society. Pancreatic Cancer Survival Rates, by Stage. (2016). www.cancer.org/cancer/pancreatic-cancer/detection-diagnosis-staging/survival-rates.html
32. Shin DW, Cho B, Guallar E. Korean National Health Insurance Database. *JAMA Intern. Med.* 176(1), 138 (2016).
33. Yeh JM, Tramontano AC, Hur C, Schrag D. Comparative effectiveness of adjuvant chemoradiotherapy after gastrectomy among older patients with gastric adenocarcinoma: a SEER-Medicare study. *Gastric Cancer* 20(5), 811–824 (2017).

34. Smalley SR, Benedetti JK, Haller DG *et al.* Updated analysis of SWOG-directed intergroup study 0116: a Phase III trial of adjuvant radiochemotherapy versus observation after curative gastric cancer resection. *J. Clin. Oncol.* 30(19), 2327–2333 (2012).
35. Cunningham D, Allum WH, Stenning SP *et al.* Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N. Engl. J. Med.* 355(1), 11–20 (2006).
36. Doria-Rose VP, Marcus PM. Death certificates provide an adequate source of cause of death information when evaluating lung cancer mortality: an example from the Mayo Lung Project. *Lung Cancer* 63(2), 295–300 (2009).
37. Penson DF, Albertsen PC, Nelson PS, Barry M, Stanford JL. Determining cause of death in prostate cancer: are death certificates valid? *J. Natl Cancer Inst.* 93(23), 1822–1823 (2001).
38. Percy C, Ries LG, Van Holten VD. The accuracy of liver cancer as the underlying cause of death on death certificates. *Public Health Rep.* 105(4), 361–367 (1990).
39. Percy C, Stanek E 3rd, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *Am. J. Public Health* 71(3), 242–250 (1981).