Published in final edited form as:
Stat Med. 2010 August 30; 29(19): 1977–1997. doi:10.1002/sim.3936.

The promise and realities of CER

Constantine Gatsonis, PhD
Center for Statistical Sciences, Brown University, Box G-S121, Providence RI, 02912

Constantine Gatsonis: gatsonis@stat.brown.edu

Abstract

Comparative Effectiveness Research has been given a broad and ambitious mandate. Will it be able to deliver the multifaceted and granular comparative information it has been tasked with developing? After a discussion of the general conditions for the feasibility of CER, we focus our attention on one of the most challenging areas: the evaluation of diagnostic tests and biomarkers.

1. Is CER methodologically feasible?

Comparative Effectiveness Research (CER) has emerged as a major initiative in the US, with broad ranging implications for both research and health care policy. As noted by Tunis and colleagues in their thought provoking article for this issue of *Statistics in Medicine*, the evidence from Comparative Effectiveness Research is intended to support *clinical and policy decision making at the individual and the population level.* Moreover, in prioritizing types of evidence, CER places a premium on the study of outcomes that are of primary relevance to patients and on the derivation of conclusions that can inform individual patient choices.

The scope of CER is broad and its mandate is undoubtedly ambitious. The new initiative attempts to overcome the fragmentation and to focus the sprawling area of biomedical research on the production and delivery of timely and comprehensive information needed to support medical decision making at all levels. It is not surprising that the magnitude of the undertaking invites a degree of skepticism. Can the biomedical research enterprise deliver such multifaceted and granular information? The answer may well be: Not without significantly changing the rules of game, that is, not without *transforming* the research infrastructure and *reforming* the overall consensus on how evidence is developed and evaluated.

Tunis and colleagues are right to devote much of their attention to these two major requirements. The *transformation* of the research infrastructure will require resources, technical innovation, and organizational flexibility. None of these are trivial matters but they are tangible and can be directly influenced by government and other institutional actors. However, the *reformation* of the consensus on evidence is rather an intangible. On this requirement for the success of CER, Tunis and colleagues can only list a set of principles and point to the need for robust methodologic development.

The first principle offered by the authors suggests that we need an alternative approach to interpreting and evaluating the strength of evidence. That much is clear. It is simply not realistic to expect that what is currently considered definitive evidence on the comparison of outcomes from two or more interventions will be generally available in a timely and detailed form that enables transparent decision making. However, the difficult question is who will determine what the new approach to evaluating evidence will be and how.

Naturally the reformation of the current consensus will involve the scientific community and its institutions and societies. I would argue that, if it came to pass, this would truly be a *paradigmatic shift*, with all the complex consequences such a shift would entail. However, it is a lot safer to pronounce a paradigm change that took place than to predict it. From today's vantage point it is more fruitful to look at the concrete modalities of CER and identify areas in need of further development. Thus few would argue with the point, made in the introduction of the paper, that a range of methodologic approaches for generating evidence are currently available, including randomized and observational primary studies, research synthesis and modeling. And more methods can be developed and undoubtedly will be in the coming years.

The second principle offered by the authors effectively restates the current lack of consensus on the evidentiary framework for CER. Surely study designs and approaches will need to be tailored to the specific subject matter questions and decision making needs. The definition of what a definitive answer might be remains elusive and is likely to be so for the foreseeable future.

As a working model of this future, I would offer that of a CER enterprise in which the research agenda is driven mainly by the decisions that need to be made, mostly about health care policy but also about the health of individuals. If *transparency* of the decision making process is desired, *quantitative* approaches will be required. Thus this perception of the future links CER closely with an area of research shunned by many in the current health policy debates: decision analysis and cost-effectiveness analysis.

2. CER for Diagnosis and Prediction

The utilization, effectiveness and cost-effectiveness of test modalities for diagnosis and prediction are already receiving considerable attention and scrutiny in the current debates about the cost and effectiveness of health care. Several of the recent IOM recommendations for CER studies relate to diagnostic tests and biomarkers either directly or as part of therapeutic interventions. Indeed, Tunis and colleagues use two diagnostic modalities, Positron Emission Tomography (PET) for cancer and Computed Tomography Angiography (CTA) as prime examples of the need for CER and of the difficulties in arriving at a consensus on what would be definitive evidence. Singling out two recently introduced imaging modalities with potentially broad applicability as examples is not accidental. Diagnostic tests and biomarkers are proliferating, can be costly, and are rather difficult to evaluate from the patient outcome perspective.

Historically, the dominant paradigm for evaluating medical care has been provided by the evaluation of therapeutic interventions. This paradigm defines what aspects of an intervention are to be evaluated at any given stage of the development of the intervention and what measures of efficacy and effectiveness are to be used. For example, the paradigm includes the well known categorization of clinical trials into four phases. The paradigm typically works with patient outcomes, such as morbidity, mortality, quality of life and functioning.

The situation with diagnostic tests and biomarkers is different. Tests generate *information*, which is only one of the inputs into the subsequent therapeutic decision making process. And most effects of the provision of diagnostic information are *mediated* by therapeutic decisions (Figure 1). Thus the assessment and comparison of the effectiveness of diagnostic tests need to account for the effects of subsequent interventions in order to identify the difference in outcome that could be reliably attributed to the test. In addition, diagnostic technology evolves very rapidly and creates a virtual *moving target* problem.

The ubiquitous role of diagnosis in health care has led to multidimensional approaches to the evaluation of diagnostic tests. The literature includes several proposals of “frameworks” for diagnostic test evaluation. Although there are differences between these proposals, they tend to be structured around three key questions:

i. How accurate is the test in its diagnostic or predictive task?

ii. Does the information from the test influence subsequent diagnostic workups and therapeutic interventions?

iii. Can the performance of the test be linked to patient level outcomes, such as those assessed in studies of therapy?

In practice, the assessment of the diagnostic or predictive accuracy of a test is typically the objective of most diagnostic test evaluations, including the comparative ones. Studies of the effect of the performance of test on subsequent clinical decision making are less common, but relatively straightforward to design and implement. For example, comparative studies in this category could be prospective, randomized clinical trials, retrospective analyses of medical records, or registries. However, studies of patient-level outcomes of tests are typically complex and resource intensive.

As noted earlier, an additional particularity of diagnostic modalities is that many of them undergo continuous technologic change. Thus, an approach to diagnostic test evaluation would also need to prescribe which of the above three dimensions of the evaluation of diagnostic tests need to be assessed at any given stage of the development of the test. Although opinion is by no means uniform, most researchers would consider diagnostic or predictive accuracy to be important at all stages of the development of a modality, with studies of process of care effects and patient outcomes being more appropriate and feasible for mature or disseminated technologies. (1)

3. Comparative studies of patient outcomes

A detailed discussion of CER for patient outcomes associated with diagnostic tests is naturally beyond the scope of this commentary. As most readers are aware, at least one class of such studies, the comparative studies of screening modalities, has received a lot of attention in the literature and has consumed large amounts of resources. However, CER studies are infrequent when it comes to patient outcomes related to the three other primary roles of diagnostic modalities in health care, namely, diagnosis and disease staging, patient management, and post-therapy surveillance. This situation will undoubtedly change in the coming years, as more attention is given to the evaluation of outcomes of tests and biomarkers, across the spectrum of health care.

The development of CER studies of patient outcomes associated with tests and biomarkers will benefit enormously from the application of the principles of an evidentiary framework for CER, presented by Tunis and colleagues. Surely a range of methods will need to be used, including but not limited to randomized clinical trials. And as surely, there is no “one-size-fits-all” approach.

When it comes to randomized studies of test outcomes, it is important to note that both long- and short-term effects of tests materialize in a context defined by the available health care options, including therapeutic interventions. It is therefore not possible to define and measure test effects outside the particular health care context in which the test will be used. For example, in the unfortunate but not altogether rare situation in which therapy offers rather ineffective options, the impact of diagnostic tests on patient outcomes will be minimal. It is also important for study designs to link test results to specific therapeutic strategies.

Even with close linkage of test results to therapeutic interventions, there are significant challenges when it comes to the practical feasibility of randomized studies of test outcomes. Consider for example a simple design for a randomized study to compare outcomes of two alternative tests, A and B, which are being evaluated for use in detecting the presence of a clinical condition. In this design, patients will be randomized to undergo test A or test B and the subsequent course of action will be based on the results of the tests. For simplicity, assume that two therapeutic interventions are available and that a positive test result on either A or B would lead to a decision to adopt the first intervention (Tx_1). A negative result on either test would lead to a decision to adopt the second intervention (Tx_2). Tx_2 could be "usual care" or active therapy depending on the context.

The simple randomized design of Fig 2 is representative of many settings in which randomized studies are conducted to compare test outcomes. Assume now that prior studies provide estimates of the success rates r_1 and r_2 for therapeutic interventions Tx_1 and Tx_2, when performed on cases that actually have the clinical condition the two tests are intended to detect. If the specificities of the two tests are equal, some algebra shows that the difference in the overall success rates between the two arms of the randomization is

$$D=(r_1-r_2)p(Sens_A-Sens_B) \quad (1)$$

where p denotes the prevalence of the clinical condition and *Sens* denotes the sensitivity of a test. As can be seen from Eq (1), even if the prevalence of the clinical condition is relatively high, the actual difference in overall success rates between the two arms of the study is only a fraction of the difference in success rates between the two therapeutic interventions.. If, as in the screening setting, the prevalence is low, the estimate of effect is even smaller.

The simplified setting of the example underscores the often stated point that randomized studies of test outcomes can be very resource intensive, to the point of becoming impractical. Insistence on the availability of randomized study results as the definitive evidence for the appropriateness of a particular diagnostic modality in a given clinical context can create the type of impasse mentioned by Tunis and colleagues in the case of PET in cancer therapy.

This is not to say that randomized studies of outcomes are to be abandoned in the case of PET or any other diagnostic modality. More efficient designs may be already available or could be developed in the future (2). However, the promise of CER is not to confine its range of possibilities within a single evidentiary dimension but to encompass a range of methodologic approaches.

An example of an approach that has been already put to practice is the use of registries. As noted by Tunis and colleagues, the National Oncology PET Registry (NOPR) provides a useful example of the possibilities and limitations of the approach (3). The registry was developed in order to assess the effect of PET on referring physicians' plans for intended patient management and to do so across a spectrum of cancer conditions in which PET may be used. To-date NOPR has gathered information on more than 130,000 cancer patients and has led to a substantial list of published reports documenting particular aspects of PET use. These reports are available in a more timely and real-world setting than would be possible in prospective clinical trials, perhaps even in the (curiously named) "pragmatic" trials advocated by Tunis and colleagues. Of course, the evidence provided by NOPR has significant limitations, including selection bias and silence on key questions about whether the intended therapy changes were actually implemented and whether they led to improved patient outcomes. Some of these limitations could be overcome by additional follow-up data collection, while others such as the potential for selection bias would apply to many types of registries.

4. Research synthesis and modeling

The particular role of diagnostic information in the process of care and the challenges of designing and conducting empirical studies of patient outcomes, provide a most appropriate setting for the effective use of modeling. Although modeling of complex health care settings has been met with skepticism by a segment of the research and policy making community in the past, the use of decision analysis and more elaborate modeling approaches, such as micro-simulation, in the evaluation of diagnostic modalities has increased rapidly in recent years. A particularly productive example of modeling research is the set of projects organized and funded by the Cancer Intervention and Surveillance Modeling Network (CISNET) of NCI (http://cisnet.cancer.gov) (4,5). Modeling methods, such as those developed by CISNET in the context of screening could be adapted relatively easily to other contexts. Such a development would facilitate the synthesis of available information from all sources in order to make credible projections about the potential impact of the use of diagnostic tests in clinical practice.

5. Conclusion

The ambitious promise of CER to provide comparative evidence suitable for individual and policy decision making about health care options may well be pointing to a distant goal. Nothing less than a paradigm shift on how evidence is developed and evaluated, accompanied by a major transformation of the biomedical research infrastructure will achieve this goal. However, a lot of progress can be made along the way. In the case of the evaluation of diagnostic tests and biomarkers, achievable and highly relevant CER studies can be developed if the full range of available methodologies is embraced, the proper role of randomized studies is understood and further enhanced, and the potential of sophisticated modeling is fully developed.

References

1. Gatsonis CA. Design of Evaluations of Imaging Technologies: Development of a paradigm. Academic Radiology. 2000; 7:681–683. [PubMed: 10987328]
2. Bossuyt P, Lijmer J, Mol B. Randomised comparison of medical tests:sometimes invalid, not always efficient. Lancet. 2000; 356:1844–47. [PubMed: 11117930]
3. Hillner B, Siegel B, Liu D, et al. Impact of Positron Emission Tomography/Computed Tomography and Positron Emission Tomography (PET) alone on expected management of patients with cancer: Initial results from the National Oncologic PET Registry. J Clin Oncol. 2008; 26:2155–2161. [PubMed: 18362365]
4. Cancer Intervention and Surveillance Modeling Network (CISNET) Breast Cancer Collaborators. The impact of mammography and adjuvant therapy on U.S. breast cancer mortality (1975–2000): Collective Results from the Cancer Intervention and Surveillance Modeling Network. J Natl Cancer Inst Monographs. 2006; 36:1–126.
5. Zauber A, Lansdorp-Vogelaar I, Knudsen A, Wilschut J, van Ballegooijen M, Kuntz K. Evaluating Test Strategies for Colorectal Cancer Screening: A Decision Analysis for the U.S. Preventive Services Task Force. Annals of Internal Medicine. 2008; 149:659–669. [PubMed: 18838717]

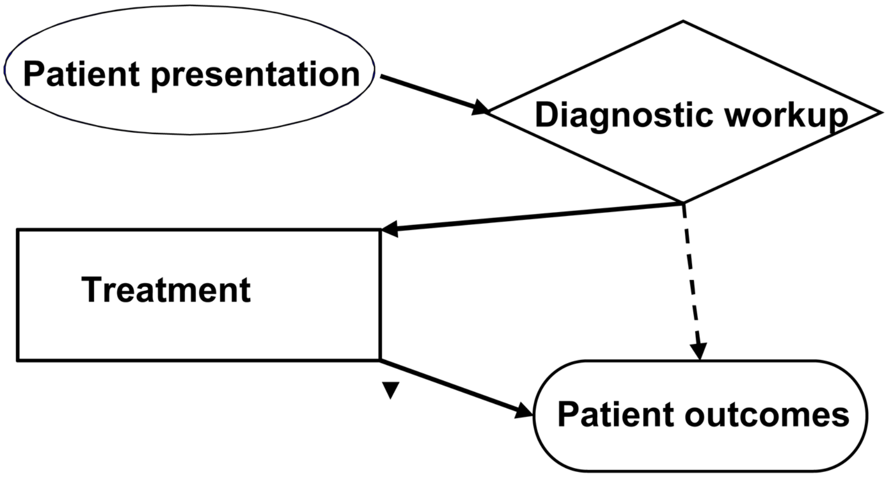

Figure 1.
Treatment mediates the effects of diagnostic testing

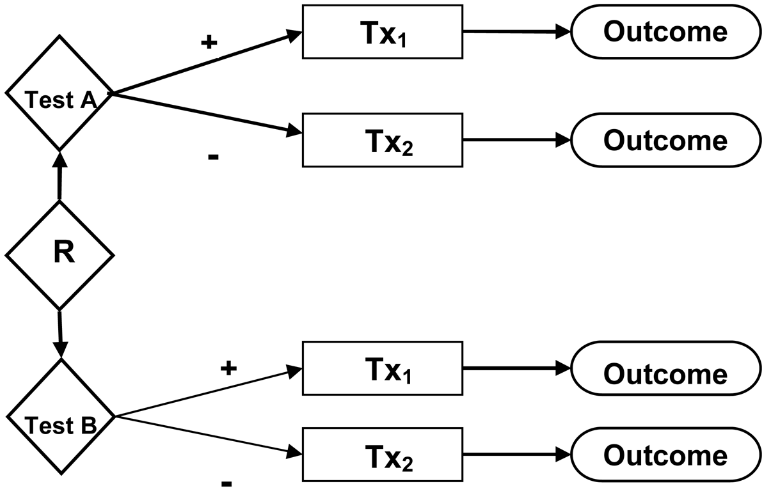

Figure 2.
Simple randomized study of test outcomes