

# A View From the Technology Evaluation Center

Margaret Piper

## Summary

The Blue Cross and Blue Shield Association's Technology Evaluation Center conducts assessments of both therapeutic and diagnostic interventions. Although each type of assessment is similar, there are unique challenges. Assessments such as these can help health plans make evidence-based coverage decisions.

## Key Points

- Health plans want to make evidence-based decisions and policies.
- There are considerable challenges to obtaining good evidence on outcomes to evaluate various therapies and technologies.
- Cost-effectiveness and affordability are becoming pressing concerns for many health plans.

TECHNOLOGY ASSESSMENT SUPPORTS health plans and other stakeholders in developing evidence-based policies. The Blue Cross and Blue Shield Association's Technology Evaluation Center (BCBS TEC) offers three products to assist their plans: medical policy, coverage policy, and payment policy manuals. These are guidelines for the various independent BCBS plans around the country, each of which uses a plan to make its coverage decisions. The technology evaluation center's purpose is to inform based on scientific evidence, and not to dictate policy.

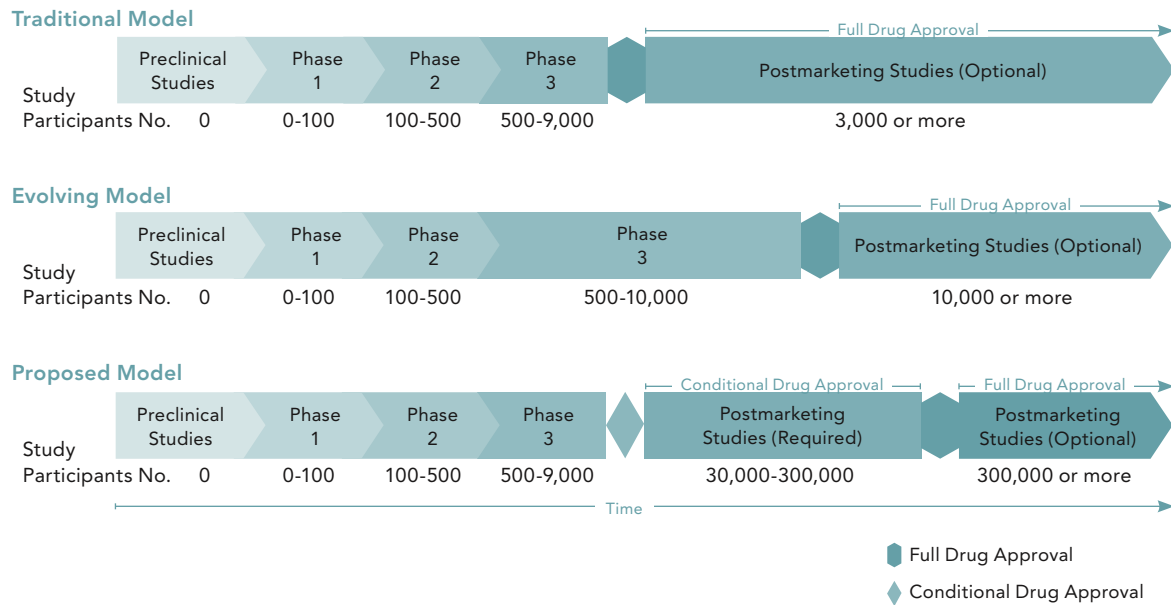
The TEC provides rigorous assessment of clinical evidence for therapeutic and diagnostic interventions through a systematic review with quality appraisal. An independent, expert medical advisory panel reviews all of the work. The panel is composed of academic clinical researchers, specialty society appointees, and a few plan representatives. The panel has the final scientific authority over each TEC assessment. The TEC culls and synthesizes the variety of evidence available pertaining to a particular topic, looking for outcomes that matter to the patient. Overall, the technology's beneficial effects on health outcomes should outweigh any harmful effects on health outcomes.

The center has conducted more than 300 technology assessments. The center maintains a three-year public inventory at [www.bcbs.com/tec](http://www.bcbs.com/tec). Some work is published in scientific journals, but that is not the focus of the center's work. The center is an Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center.

One of the first steps in a technology assessment is identifying available studies on the topic. Good clinical studies are like good accounting and financial practices. Deviation from standards leads to unreliable results and misinformed decisions. Technology assessments are trying to determine if the benefits of a technology outweigh the costs and risks, so good quality studies are given priority in the assessment process.

In performing the technology assessments, there are many challenges in assessing outcomes. These include the quality and quantity of trials; inconsistent reporting of adverse effects; the lack of robust evidence of effects and comparative effects; and selective reporting and publication bias. Randomized controlled trials are the gold standard but are not always available. The center tries to determine whether a topic's benefits can be determined by other methods such as observational data.

**Exhibit 1: Alternative Models for Studying Drug Safety**



Adapted from reference 3

The ideal outcome measure is one that the patient can perceive. Unfortunately, for many topics, this ideal outcome would require large, very long-term, and expensive studies, which are difficult to have funded. Intermediate outcomes are commonly used. These can be measured at a shorter time period or more easily than the ideal outcome, but are strongly related to the final outcome of interest. Measurement of outcomes that have a high degree of subjectivity, such as pain, have to be closely defined, and the clinical significance of the differences needs to be defined.

For example, there are a lot of quality-of-life scales that have been developed for general health assessment or particular diseases. These scales distill quality of life into some numeric result. The question is whether the statistical difference in numeric results is really clinically significant. These scales need to be validated. Unpublished and non-validated scales tend to show larger effect than published, validated scales. Some studies will report a composite outcome, but these can be misleading. The composite outcome may be driven by the least important outcome. An example is the rate of TIA (transient ischaemic attack) driving the total instead of stroke.

Surrogate outcomes are used in many studies. A surrogate outcome known to be related to the final desired outcome would be ideal. The surrogate outcome should be in the causal disease pathway. Validation of a surrogate outcome shows whether it lies in the causal pathway.

There can be problems using surrogate outcomes. Diseases can have multiple causal pathways. Additionally, surrogate outcomes may not identify unintended adverse effects of the intervention being studied. Surrogate outcomes are commonly used to speed up clinical trials when the final outcome would take years to identify, and to reduce costs.

Overall, in assessing an intervention, first in the hierarchy of outcomes is a true health outcome that matters to the patient. Second is a surrogate that is validated to reliably predict the final outcome. Third is a unvalidated surrogate that is reasonably likely to predict the final outcome. Last in the hierarchy is an outcome that correlates with biological activity but may not predict the final outcome.<sup>1</sup>

Another problem area with therapeutic intervention assessments is establishing the true adverse effects of a particular medication. There can be problems with adverse effect reporting in clinical trials. When studies measure adverse effects, there is variation in the precision of definitions; what is reported; and how events are classified into categories. This makes it difficult to compare across trials. Withdrawal from therapy because of adverse events is a common summary measure, but why subjects withdrew is often not defined. It is not always enough to know how frequently an event occurs, but it is also important to know the severity of effects. In an attempt to improve adverse event reporting in certain trials, the Radiation Therapy Oncology Group has developed criteria to grade

**Exhibit 2: Diagnostic Model  
A Continuum for Efficacy**

	<u>Paraphrased</u>
• Level 1: Technical efficacy	Pretty Picture
• Level 2: Diagnostic accuracy efficacy	Improved Accuracy
• Level 3: Diagnostic thinking efficacy	Improved Diagnosis
• Level 4: Therapeutic efficacy	Improved Treatment
• Level 5: Patient outcome efficacy	Improved Health
• Level 6: Societal efficacy	Improved Efficiency

Copyright 2008 Blue Cross Blue Shield Association

toxicity severity. All participants in this research group have agreed to use the same criteria.

Exhibit 1 illustrates the differences between the traditional model, the evolving model, and a proposed model of evaluating adverse effects from medications.<sup>3</sup> The difference in these models is largely between phase III and post-marketing studies. Often at approval, the Food and Drug Administration asks companies to conduct post-marketing studies, but these studies may or may not get done. The evolving model, which is happening now, is to widen phase III trial process to gather more information before full drug approval. The consequences of the evolving model are longer and more expensive trials, and a delay of getting products to market. The proposed model is to maintain the shorter phase III period; review data; give conditional approval if appropriate; and give final approval when a post-marketing study is completed, and appropriate risk versus benefit has been demonstrated. Additional adverse event studies might be suggested at the time of final approval.

Overall, robust evidence of effects requires high-quality trials with long-term follow-up to assess benefits and harms. Surrogate outcomes may be misleading. Comparative trials may be needed to address clinical questions rather than regulatory issues.

Some studies have suggested a selective reporting and publication bias. Negative or insignificant results and adverse events have been of less interest to publish. Although industry-sponsored trials can be rigorously designed and conducted, there are questions about the relationship of sponsors to outcomes and conclusions.

Diagnostic technologies have to be evaluated in a similar manner to new medications. A diagnostic technology would be a laboratory or imaging test that is used for disease diagnosis, risk, or prognosis. There is a continuum of efficacy with diagnostic

technology (Exhibit 2). Whether the test provides an accurate result is level 1, technical efficacy. Beyond technical efficacy, the clinical world wants to know how well a test's results relate to the clinical outcome of interest – clinical validity and utility. Lastly, level 6 assesses whether a test has a large public health impact. A model process for evaluating data on emerging genetic tests is shown in Exhibit 3.<sup>4</sup> This model incorporates analytical validity, clinical validity, clinical utility, and societal impact.

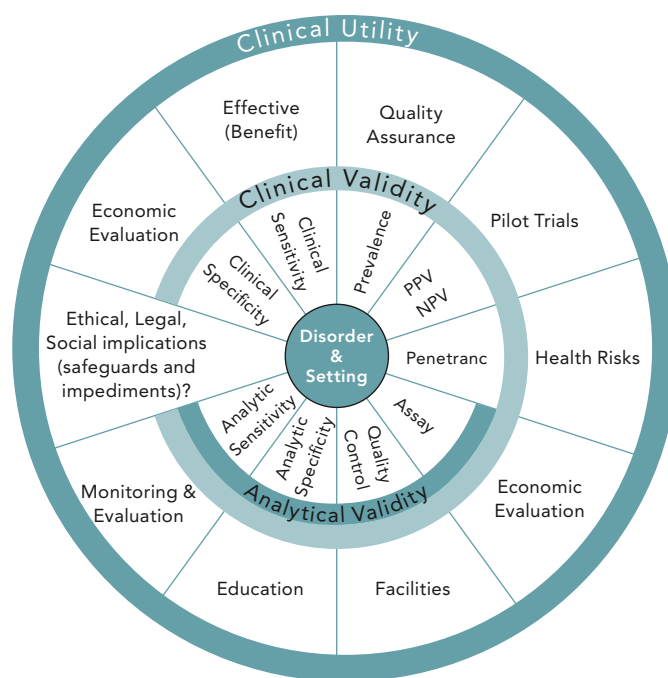
Ideally, there would be randomized controlled trials that compare the new test or procedure to standard care without this test. Particularly in the diagnostic arena, this is difficult. Typically, the diagnostic industry is not funded as the pharmaceutical industry is for conducting large-scale, long-term clinical trials. Money for translational research (i.e., from basic to clinical settings) has largely been overlooked by government agencies and other funding sources. Because of these issues, other options, such as indirect evidence, have to be used. This involves constructing a causal diagnosis-treatment chain such that observational evidence is used to translate the likely impact on health outcomes as a result of using the given test. If there is a disruption in the causal chain, this can be difficult to accomplish.

As with treatment technologies, study quality is evaluated – the study population, description of the test, the reference standard used, and any blinding are examples of what would be evaluated. As an example, the population of the final validation study for a test should be representative of the intended population for a given test. The best way to do this is with a costly, time-consuming prospective study. The ultimate question when evaluating a diagnostic test is whether the test offers an incremental benefit over what is already available.

Another issue is the public availability of information on laboratory tests being offered without FDA clearance, but in a CLIA-licensed laboratory. Many times information on test performance or validity is not available. This is particularly a problem with many of the new genomic tests.

One example is the genetic test for long QT syndrome. Several genetic markers have been discovered that predict long QT, which in some cases can lead to sudden death. The genetic test is used to confirm the syndrome when it is suspected in a patient, and to screen relatives of patients with known long QT syndrome. Once identified as having the syndrome by the test, the patient can be treated with beta-blockers to prevent complications. The evidence for the effect of

Exhibit 3: The ACCE Evaluation Process for Genetic Testing



Copyright 2008 Blue Cross Blue Shield Association  
Reference: 4

beta-blockers is incomplete, but it is a low-risk intervention. Although all the information is not available, the use of the test and the use of beta-blockers meet standards for use.

Another example is genetic testing for markers that predict warfarin dosing. Testing can predict the final dose rather than have the patient started on standard dosing requiring adjustments. It is possible that this will decrease the time to stable anticoagulation and possibly complications. The problem with this type of testing is that genetics is not the only influence on stable warfarin dosing. In European Caucasians, genetic variants account for one-third to one-half of the variability in stable warfarin dose. At this time there is not enough information that this test predicts stable warfarin dose with enough accuracy that the patient will be in the treatment window sooner with improved outcomes. Several randomized controlled trials are underway assessing this test.

BCBS TEC assessments do not typically include costs. Cost-effectiveness is not yet a driver of medical policy. Currently, clinical effectiveness is the driver. Formal cost-effectiveness analysis is a better indicator than simply comparing the cost of a test against the benefit.

New technologies may bring small benefit at high cost. As more of these costly technologies

reach the market, health care costs are increasing astronomically. Cost-effectiveness and affordability are pressing issues for many health care plans. Plans have to assess how much cost they can afford balanced against a small benefit. Employers also are concerned about these costs.

### Conclusion

Health plans want to make evidence-based decisions, but there are considerable challenges in obtaining good evidence on outcomes. Evidence-based decisions can be made for therapeutic and diagnostic interventions. More and more, cost-effectiveness and affordability are pressing concerns for health care plans. **JMCM**

**Margaret Piper** is director of genomics research, Technology Evaluation Center, for the Blue Cross and Blue Shield Association.

### References

1. Fleming, T. Surrogate endpoints and FDA's accelerated approval process. *Health Affairs*. 2005;21:67-78.
2. Ledford H. Weighing up the evidence. *Nature*. 2007;447:512.
3. Kuehn BM. IOM: overhaul drug safety monitoring. *JAMA*. 2006;296:2075-6.
4. Haddow JE, Palomaki GE. ACCE: A model process for evaluating data on emerging genetic tests. In *Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. Khoury M, Little J, Burke W (eds.), Oxford University Press, pp. 217-233, 2003.