

1-24-07

Health Technology Assessment and Comparative Effectiveness:
Recommendations for Improving Health Care Value in the United States

Prepared for FRESH-Thinking:
Focused Research on Efficient, Secure Healthcare

Steven D. Pearson, MD, MSc, FRCP
Division of Clinical Bioethics, National Institutes of Health
Institute for Clinical and Economic Review, Harvard Medical School

I. Introduction

The United States spends nearly twice as much per capita on health care as any other nation. Moreover, for the past 20 years the US has sported one of the fastest growth rates in health spending among developed countries. Health economists and policy experts believe that the adoption of new technology --- including new drugs, devices, procedures, and biologics --- plays a major role in sustaining these trends. Yet despite their importance new technologies are often adopted and become widely used in the absence of sound medical evidence on their benefits, harms, and costs compared to established alternatives. As a result, the quality and affordability of health care in the United States have suffered. Analysis of the US health care system reveals wide regional variations in treatments, an unacceptable number of preventable medical errors, significant underuse of recommended “best practices,” and the inappropriate use of services that yield little or no demonstrable value. Policy makers and proponents of health care system reform, including those from diverse points on the political spectrum, have thus concluded that cost control and improved quality within the health care system will require more explicit appraisal of the clinical effectiveness and comparative value of new technologies.

While concern with the lack of good information on new technologies is not new, three recent developments have brought renewed attention to the evidence gaps that plague the health care system. First, after several years of relatively slow growth in the 1990’s, the cost of health care, both in the private and public sector, has resumed a pattern of rapid growth beyond the core growth rate of the US economy, creating a vision of a “fiscal hurricane” on the horizon as the US population continues to age. The rapid rise in costs is threatening to tear apart the already frayed system of employer-based health insurance. Insurers and employers have already acted to shift more financial risk for the cost of care to patients through deductibles, co-insurance, and various types of reductions to the scope of insurance benefits. Insurance benefit innovations in the future will require better information on benefits, harms, and costs of technologies, information needed equally by patients who are assuming greater responsibility for the costs of care.

Compounding concern about rising costs is anticipation of the impending wave of new therapies based on genetic and other complex biologic mechanisms. These therapies herald an era of “personalized” medicine, one in which therapies can be targeted more precisely for individual patients, enhancing the rates of success while minimizing side effects. But many of these new treatments are being introduced at prices much higher than those of previous new drugs. Costs in excess of \$50,000 - \$100,000 for a course of treatment are now becoming routine. Spending on these new “specialty pharmaceuticals” reached \$42 billion in 2004 and was estimated to have risen to \$69 billion in 2006, representing 25% of the nation’s pharmacy bill.

The third recent development spurring interest in better assessment of technologies is the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. The MMA obligated the federal government to provide drug coverage for seniors at a cost of hundreds of billions of dollars over the next decade. The MMA also explicitly prohibited the federal government from using the full weight of Medicare’s purchasing power to

drive down prescription drug prices. This prohibition has heightened awareness among policy makers of the importance --- and current lack --- of information on the relative clinical effectiveness and cost-effectiveness of alternative therapies, one of the few remaining tools that might be used to moderate the costs of the Medicare drug benefit.

The confluence of these trends and events has recently led to numerous calls to expand the capacity for “comparative effectiveness” research in the United States. In academic articles, conference presentations, testimony before Congress, and newspaper editorials, better information on the comparative effectiveness of new technologies has been touted as being able to promote a more quality-focused, cost-effective health care system. Comparative effectiveness information is viewed as critical, not only to inform decision-making by insurers but also to help physicians and patients make more informed, and presumably, more prudent decisions regarding the use of new technologies.

It is important here to note a distinction, and a relationship, between “technology assessment” and “comparative effectiveness.” The evaluation of specific health care technologies has traditionally been called technology assessment, but in the recent policy debates in the United States the broader term comparative effectiveness has dominated, supplanting after a long decline its antecedent, “outcomes research.” Comparative effectiveness is most usefully viewed as a broader concept within which can be placed two specific functions: 1) Evidence development – clinical trials and other types of research intended to generate new evidence; and 2) Technology assessment – the systematic review and analysis of existing evidence, sometimes supplemented with new meta-analyses and decision analyses, to help inform health care decisions. Viewed this way, technology assessment figures as one component within the scope of a broader comparative effectiveness initiative.

Although this paper will discuss the ideal relationship between these two components of comparative effectiveness, the primary purpose will be to focus on the structure and methods of technology assessment in the United States in order to propose a set of recommendations to address the existing barriers that have limited its usefulness in efforts to improve the value of health care. The paper will first review the current status of technology assessment in the United States, highlighting specific structural and methodological challenges. Second, an overview of relevant international examples of technology assessment initiatives will focus on insights from the experience of other nations’ attempts to integrate technology assessment within a broader program to improve the value of care. The paper will then turn to a series of specific recommendations for ways that technology assessment can be enhanced in the United States. Emphasis will be given to methodological innovations that can strengthen the rigor, legitimacy, and usefulness of technology assessment within the US health care system. A concluding section will describe the scope and mission of a new federal role in coordinating technology assessment as part of a broader comparative effectiveness initiative. Linking technology assessment to the function of generating new evidence on comparative effectiveness will be described as the best way for technology assessment to play its fullest and most useful role in improving the quality and efficiency of the US health care system.

II. Technology Assessment Efforts in the United States

More than 1,000 public and private sector organizations in the United States conduct technology assessments, including Federal and State agencies, non-profit organizations, health plans, professional societies, for-profit consulting companies, health industry manufacturers, and hospitals. With no dominant federal involvement, the picture of health technology assessment in the United States is thus marked by its independence and diversity. In the section below the goal will be to describe the main outlines of technology assessment at three levels: the federal government, the states, and within the private sector. The most influential examples, both structurally and methodologically, will be highlighted in order to elucidate the key opportunities to strengthen the role of technology assessment.

Federal health technology assessment:

The federal government's first major HTA effort began with Congressional passage of the Food Drug and Cosmetics Act in 1938, which established many of the powers that define the modern Food and Drug Administration (FDA). The FDA's original mandate to review drug safety was expanded to include clinical efficacy in 1962 and, in 1976, Congress extended the FDA's regulatory authority to include the evaluation of medical devices.

Clinical trials sponsored by manufacturers are designed largely to meet the needs of the FDA approval process, but the evidence generated by these trials is often insufficient to guide decisions by insurers, physicians, or patients. The FDA evaluates only the safety and efficacy of new therapeutic agents, and usually makes its determination based on studies of relatively small, targeted patient groups. Health care decision-makers want to know the effectiveness of new technologies in real practice, in broad populations of representative patients, and in comparison with other established therapies. But providing this information is not part of the FDA's mission, nor does the FDA consider in any way the costs or cost-effectiveness of new technologies. Without these broader considerations, the FDA regulatory review process cannot by itself support value-oriented decision-making in the health care system.

Although the FDA has never been involved in technology assessment beyond questions of safety and efficacy, beginning in the 1970s there have been several federal technology assessment initiatives with a broader mandate. These efforts included the Congressional Office of Technology Assessment (OTA), the National Center for Health Care Technology (NCHCT), the Office of Health Care Technology Assessment (OHCTA), and the Agency for Healthcare Policy and Research (now AHRQ). Of these organizations only AHRQ remains. A full history of the rise, fall, and – for AHRQ – near fall of these various initiatives is beyond the scope of this paper; but the demise of each organization was framed by the absence of a strong constituency in the face of intense opposition from groups like the American Medical Association and the Health Industry Manufacturers Association.

While other federal health agencies, including the Centers for Disease Control and the Veterans Administration, engage in health technology assessment, a chastened AHRQ remains today as the flagship agency performing health technology assessment at the federal level. In addition to several legacy programs, AHRQ received a new impetus and direction for technology assessment as part of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. Section 1013 of the MMA calls on AHRQ to conduct research on the "outcomes, comparative clinical effectiveness, and appropriateness of health care, including prescription drugs." Under this initiative, AHRQ has expanded its work conducting comparative effectiveness reviews of drugs and other technologies. The initiative shows the potential of the federal government to play a more active role, although, to date, the amounts authorized for the effort have been small, on the order of \$15 million per year. In addition, as a tangible sign of the ambivalence toward federal research on comparative effectiveness, MMA prohibits the Centers for Medicare and Medicaid Services (CMS) from actually using the evidence from AHRQ reviews to withhold coverage of a prescription drug under Part D of the Medicare benefit.

AHRQ has several other, partly overlapping programs that sponsor technology assessments of one kind or another. First, despite the prohibition in the MMA, AHRQ has a longstanding arrangement to provide technology assessments for CMS. These technology assessments, focused on procedures and diagnostic devices, are used by CMS to inform its national coverage decisions for the Medicare program as well as provide information to Medicare carriers. Many of these technology assessments are done in collaboration with one of AHRQ's Evidence-based Practice Centers (EPCs). The EPCs, now numbering 12, were initially launched in 1997 to promote evidence-based practice in everyday care. Their mission is to develop evidence reports and technology assessments on topics relevant to clinical, social science/behavioral, economic, and other health care organization and delivery issues—specifically those that are common, expensive, and/or significant for the Medicare and Medicaid populations. Topics for EPC assessments are nominated by non-federal partners such as professional societies, health plans, insurers, employers, and patient groups.

Separately, AHRQ also provides core funding for a network of Centers for Education and Research on Therapeutics (CERTs). The CERTs is a national initiative to conduct research and provide education that advances the optimal use of therapeutics. Some of this work has included technology assessment. The CERTs began as a demonstration program authorized by Congress as part of the Food and Drug Administration Modernization Act of 1997 and the full CERTs program was established as part of the Healthcare Research and Quality Act of 1999. By September 2000, AHRQ had funded a total of seven centers, with each center focusing on therapies used in a particular patient population or therapeutic area.

As part of its broad agenda in technology assessment, AHRQ has also worked through several of its programs to fund efforts to improve the methods of evidence-based systematic review, the process that is at the heart of technology assessment. This work on methodology has included workshops and papers on statistical methods, on literature review and summary, and on the grading of evidence. In addition, even though AHRQ has only rarely included cost-effectiveness in any of its technology assessments, it has

sponsored work seeking to improve the consistency and rigor of cost-effectiveness methods.

Technology assessment efforts at the state level:

Each state must have some internal process for assessing new health care technologies as part of the management of its Medicaid program. One particular state initiative has caught significant attention for its breadth, strong grounding in evidence-based methods, and controversial application to the design of Medicaid formularies. The Drug Effectiveness Review Project (DERP) is an alliance of fifteen states and two private organizations which have decided to pool resources to synthesize and judge clinical evidence for drug-class reviews. The outlines of DERP first began to emerge in 2001, when researchers at the AHRQ-sponsored EPC at Oregon Health and Science University first began conducting reviews of therapeutic drug classes for the Oregon Medicaid program. Officials in Idaho and Washington State soon began drawing upon these reviews. In 2003 the DERP was formally launched as a vehicle to invite other states and private organizations to help shape and use drug-class reviews and to share project costs. The DERP commenced its reviews in November 2003, with ten member organizations, and has since expanded to include seventeen participants: fifteen states (Alaska, Arkansas, California, Idaho, Kansas, Michigan, Minnesota, Missouri, Montana, New York, North Carolina, Oregon, Washington, Wisconsin, and Wyoming) and two non-profit organizations (the California HealthCare Foundation and the Canadian Agency for Drugs and Technologies in Health).

The DERP has not included cost-effectiveness as an element of its reviews, opting instead to judge the relative effectiveness of drugs that are considered to be within the same drug class and therefore are assumed to have largely the same therapeutic intent and eligible patient populations. DERP's assessments have aroused the animosity of some drug manufacturers and patient advocacy groups. Criticism has been leveled at the restriction on the types of evidence that will be considered by DERP in its reviews. Stakeholders also complain that the framework for DERP evaluation places too high a burden of proof for new drugs to be considered as superior to existing alternatives, resulting in an unfair "default" position that all drugs are created equal until proven otherwise. The greatest concerns, however, have focused on deficiencies in stakeholder involvement and transparency in the review process and, ultimately, on what is perceived to be the blunt application of DERP reviews by state decision makers who, critics claim, have an overriding agenda of using DERP reviews to cut costs. Whether this final claim is true or not is hard to determine, but there is evidence of inconsistent use of the reviews by state decision makers. A recent Henry J. Kaiser Family Foundation report found important differences in how four state Medicaid programs use DERP reports. Some states (such as Washington) use reports as the main source of clinical evidence for formulary development, while others (such as North Carolina) use them as one of many inputs. Participating organizations have used DERP reports not only for Medicaid coverage decisions but also to inform drug coverage policy for state employees or other public programs. Notably, Consumers Union and AARP, although not DERP members, have begun adapting DERP reviews for consumers.

Despite the criticism and pressure from some stakeholders, the reach of the DERP appears to be growing. The DERP has been reviewing or re-reviewing twenty-six drug classes over a three-year span ending September 2006. Other states and organizations are considering participation, and DERP officials are also discussing the option of conducting additional reviews, including class-versus-class reviews for selected conditions and reviews of clinical guidelines. DERP is something of a political innovation: a non-governmental entity to conduct drug reviews for mostly government clients. It thus represents an important and growing initiative, seen by many as a model for a growing role for government in the coordination of technology assessment.

Health technology assessment in the private sector:

Health technology assessment is conducted or sponsored by many private health care institutions. For many years drug and device manufacturers have conducted extensive technology assessments themselves in order to gauge the potential of new technologies and to position their applications for regulatory and coverage approval. The capacity for and sophistication of technology assessments by manufacturers has increased as governmental bodies in international markets have added requirements for specific evidence of comparative effectiveness and cost-effectiveness as part of an application for funding within their health care system.

Private health technology assessment companies in the United States, including ECRI, Hayes, Inc., Cerner, Innovus, and many others, provide technology assessment services to clients from across the spectrum of health care. These companies produce assessments that differ widely in scope and methods; products are tailored to meet the varying needs of individual clients. Among the most active consumers of these technology assessments are private health plans. Private health plans often purchase technology assessments as an element of their own internal coverage decision-making process. National health plans, such as Aetna, UnitedHealthcare, and Wellpoint, have large and sophisticated teams dedicated to technology assessment as part of setting national medical coverage policies. Smaller health plans may command fewer resources, but due to anti-trust provisions even the smallest health plan must perform its own technology assessment when a new technology comes into use to determine whether it meets the contractual standard for coverage --- that it is “medically necessary.”

Medical necessity has long eluded easy definition, and the linkage between technology assessment and coverage decision-making is therefore fraught with difficulty. One of the most influential approaches to link technology assessment to coverage was pioneered by the Technology Evaluation Center (TEC), established by the Blue Cross and Blue Shield Association in 1985, and now administered in partnership with Kaiser Permanente. TEC assessments focus on clinical effectiveness and appropriateness within specific patient populations, but do not generally consider costs or cost-effectiveness. The framework that TEC uses to assess technologies has adapted or fully adopted by many other technology assessment initiatives in the United States and so is worth exploring in some

detail. The TEC uses five formal evaluation criteria for judging the effectiveness of a medical technology. These five criteria are:

- The technology must have the final approval from the appropriate government regulatory bodies, if applicable.
- The scientific evidence must permit conclusions concerning the effects of the technology on health outcomes.
- The technology must improve the net health outcome.
- The technology must be as beneficial as any established alternatives.
- The improvement must be attainable outside the investigational setting.

TEC uses a formal approach when reviewing the evidence and judging a technology against each of these five criteria. A technology assessment is first prepared by the core staff of the TEC and then presented to its Medical Advisory Panel, or MAP. The MAP is composed of nationally respected clinical and methodological experts, with membership diversified to ensure representation by a wide variety of viewpoints within the healthcare community. A majority of members hold academic appointments and are independent medical experts without affiliation to healthcare payers.

The MAP judges a technology separately upon each of the five TEC criteria. These five criteria, which have served TEC and the health plan community well for over 20 years, were developed primarily to help private health plans judge whether a new technology was no longer investigational or experimental. This has been a key distinction and threshold for coverage decisions, since health plan contracts generally exclude coverage for all experimental and investigational services. With little support from court decisions for other evidence-based methods as a basis for judging medical necessity, disqualification as investigational or experimental has been the mainstay of the use of technology assessments by private health plans. Thus the TEC criteria were designed primarily to help provide the basis for a dichotomous “yes/no” coverage decision on new technologies and this approach continues to be the most prominent and respected approach in the United States to informing coverage decisions.

III. Limitations of Technology Assessment Efforts in the United States

From the brief summary above, four broad features of technology assessment in the United States can be identified as presenting the greatest barriers to improving the value of health care. These features are 1) Poor coordination; 2) Weak legitimacy; 3) Limited usefulness; and 4) Incomplete integration. In some ways these four deficits overlap and reinforce each other, but their distinct contributions to the limitations of technology assessment will be examined in turn.

Poor Coordination

The conduct of technology assessment is diffused across many sectors of the health care system. In stark contrast to several international examples, the United States does not sponsor a dominant centralized or coordinating technology assessment organization. Although AHRQ is nominally charged with the conduct of comparative effectiveness research, it has received scant resources and operates within a restricted mandate. In some ways the rise of the DERP, and the ongoing activity of so many private technology assessment companies, is an environmental response to the growing need for evidence on new technologies in the absence of a strong federal role.

It can be argued that diffusion, and the resulting diversity, is a beneficial feature of technology assessment in the United States. In this view diversity provides a competitive market, constantly balancing products to the needs of consumers, and ensuring choice of methods and organizations. Even when assessments of some technologies by different entities arrive at varying conclusions, this variation is useful as a means to explore the nuances of technology assessment and can serve as a type of quality assurance.

Diversity and competition may in fact have some salutary effects, but a broader view suggests that greater coordination would augment the ability of technology assessment to improve the value of care in the US. First, the lack of coordination results in duplication and an inefficient use of limited assessment resources. It is not unusual for some new technologies to be evaluated multiple times in overlapping years by different groups, both at the federal, state, and private level. The amount of resources spent on these often redundant efforts could be harnessed to expand the number of technologies assessed. But with no single authoritative body to coordinate health technology assessment, no “trusted resource,” decision-makers are often left with no recourse but to commission or conduct their own assessments.

Multiple technology assessments present a related problem: quality control. While diversity has welcomed the excellence and rigor of some technology assessment programs, such as the TEC, it has also made room for many assessments of poor quality. Instead of serving to highlight excellent methods, wide variation in methods has cast doubt on the reliability and validity of the evidence-based methods that underlie technology assessment. Nomenclature, definitions, statistical approaches, and other features of technology assessments differ widely. AHRQ has tried but has not had the resources or prestige to serve as a standard-setting body for technology assessment. The DERP also represents an attempt to coordinate assessments in order to achieve superior efficiency and technical quality in assessments, but like AHRQ it does not have the influence or authority to capture the methods of other technology assessment efforts in its orbit. Poor coordination thus often leaves important decision makers working on their own without the ability to judge assessment quality. Confusion with the multiplicity of formats and frameworks used to assess evidence is a natural outcome. The lack of standardization of methods also offers an easy target for critics who are unhappy with the result when their favored technology receives a less than stellar review.

Weak Legitimacy

The inadequacies of methodological consistency and rigor that are fostered by a lack of coordination represent one of the elements that undermine the legitimacy of technology assessments in the United States. Weak legitimacy also flows from clear deficits in objectivity, stakeholder involvement, and transparency, deficits that characterize most health technology assessment efforts in this country.

Objectivity is critical. Given that technology assessment involves judgments about the strength of medical evidence, an independent, objective perspective is necessary in order to command the respect of all important stakeholders. Much technology assessment, however, is sponsored either by manufacturers, by physician groups, or by private payers, all of whom have obvious vested interests in the outcome of any assessment. It is far too easy to question the integrity of a technology assessment performed by a stakeholder such as a manufacturer; equally doubtful often is the objectivity of technology assessments purchased from private companies by insurers. The work of DERP and TEC has been questioned by some as too closely linked to the interests of a particular stakeholder. By contrast, the technology assessments performed by AHRQ have the greatest independence from specific vested interests. But because costs are not explicitly considered as part of the technology assessment program at AHRQ, it is known from past experience that some stakeholders may still question whether the federal government harbors cost containment as a hidden agenda, and this suspicion by itself weakens the perceived objectivity and undermines the legitimacy of AHRQ's work.

Allowing stakeholders to have robust engagement with the technology assessment process is another key element of legitimacy with which the US system struggles. In general, because objectivity and freedom from conflict of interest is often viewed as the greatest challenge for technology assessment programs, stakeholders have been consciously excluded from participating in ways that might offer them too much influence. Thus neither the programs at AHRQ, the DERP, nor the TEC offer a formal partnership role to manufacturers or provider groups, outside of allowing them to submit evidence that may, or may not, be used in the assessment. The technology assessments performed by public and private insurers do not usually engage manufacturers, and insurers have even become skeptical about participation of patient groups now that evidence has emerged on the financial backing of many such groups by manufacturers. The unresolved tension between efforts to insure objectivity and those that would open the process more to stakeholder involvement is a serious problem that contributes significantly to questions about the legitimacy of technology assessment efforts.

Transparency is often linked to stakeholder participation conceptually because transparency can compensate for restrictions on participation that must be taken to insure the objectivity of a technology assessment process. Transparency can also help confirm the objectivity of the process. Unfortunately, transparency is often another weakness of technology assessment in the US. Some of the technology assessment programs within private insurers have long been criticized as being "black boxes," proprietary mechanisms kept secret even from physicians and patients directly affected by decisions. Recent years have seen significant progress as health plans and other decision-making

bodies have made available on the internet full accounts of their technology assessments. Other assessment efforts, including those at AHRQ, DERP, and TEC, have long been noted for providing detailed descriptions of the methods used to assess the evidence in their assessments. Yet even these programs could benefit from a more standardized approach to describe how they consider the relative strength of different bodies of evidence. Several national and international groups are currently working on systems that will make the judgments of evidentiary strength more reliable and transparent to all stakeholders.

Even with improved descriptions of the process of evidentiary review at the heart of technology assessment, there remains the question of how to make more transparent the ultimate decision-making process for which technology assessments are only one ingredient. Little conceptual work has been done to describe how evidence is viewed within a context of ethical, legal, financial, and other considerations. In a room decision-makers sit, each with a well-done technology assessment before them. How does a decision about reimbursement, or coverage with certain restrictions, or placement on the second tier of a formulary, emerge? Unfortunately, while our ability to describe the structure and hierarchies of evidence-based medicine has improved, evidence-based decision-making often remains stubbornly opaque to many stakeholders. The consideration of cost is particularly important in this regard. All stakeholders would agree that costs should and do play some factor in coverage and reimbursement decisions, but technology assessments usually avoid costs as an explicit component, and decision-makers often have no framework within which costs can be explicitly and legitimately discussed. This disconnect between the realities of health care decision-makers attempts to improve value, and the tools that current technology assessment programs give them to do so, is a major theme to which this paper now turns.

Limited Usefulness

There is no doubt that technology assessments have broad applicability throughout the health care system. They are used to inform decisions including investment in new technologies, strategies for regulatory approval, insurance coverage, reimbursement, medical management, and the treatment of individual patients. Yet there are several important ways in which the current structure and methods of technology assessments in the United States severely limit their application as tools to improve the value of health care.

First, there is a recognized but largely unaddressed tension between the rigor of the technology assessment process and the timeliness of the information for decision-makers. Private health technology assessment companies are acutely aware that their products must meet the timelines of decisions being contemplated by their customers, including health plans which face great pressures to render coverage decisions soon after the introduction of a new technology. But the trade-off for timeliness is too often the thoroughness and rigor of the assessment. In contrast, the assessments produced by AHRQ and its academic EPCs, often impeccable, are produced on an academic time scale notoriously disconnected from the needs of decision-makers, often taking six

months or longer to complete. A clear and consistent approach to make rigorous methods applicable within reasonable timelines is badly needed to help improve the usefulness of technology assessments.

Another limitation to the usefulness of technology assessments results from the failure to adequately combine systematic review and decision analytic modeling as complementary evidence-based methods. Commentators have noted that technology assessment in the United States is dominated by systematic review, whereas Europeans seem more comfortable with decision analytic modeling. The scope of this paper does not allow for a full discussion of the relative merits of these two approaches to evidence synthesis and analysis, but in general it can be said that decision analytic modeling has particular usefulness in situations when systematic review may be of limited use to decision makers. For example, when patients' perspectives, values, and utilities are particularly important in judging the outcomes of care, decision analytic modeling is uniquely suited to providing a framework for judging the value of care for different types of patients. Decision analytic modeling also is particularly useful when there is no direct head-to-head comparative evidence between two alternative treatments. In such situations, adherence to traditional evidentiary hierarchies might lead a systematic review to conclude that there is "inconclusive" or "inadequate" evidence to make a comparative judgment between two alternative therapies; but decision analytic modeling can take the a broader range of available evidence, supplement it where necessary with consensus or expert advice, and present a model through which the effectiveness of two treatments can be compared. Decision analytic modeling accepts a greater degree of uncertainty in the evidence and creates a specific framework within which to assess how that uncertainty may affect what can be judged about the comparative effectiveness of a technology. Modeling has its own set of acknowledged vulnerabilities, and usually adds complexity and time to any technology assessment, but, overall, inclusion of modeling as a standard approach within technology assessment programs in the United States would likely improve the usefulness of the results for many decision-makers.

Decision modeling also is the platform on which cost-effectiveness analyses may be built, which introduces another key limitation to the usefulness of many technology assessment programs: the failure to grapple with comparative value as a component of comparative effectiveness. The separation of economic evidence from clinical evidence is understandable as a political construct, given that open consideration of cost-effectiveness remains rare in the United States. AHRQ, DERP, and TEC all have adopted and strongly reinforce the convention among technology assessment organizations to consider clinical evidence on its own merits, without respect to costs. However, the lack of procedures for considering economic evidence more explicitly creates its own set of problems. As mentioned, it likely contributes to some of the distrust of technology assessment efforts, since stakeholders assume that costs are being weighed in some surreptitious manner during the review of evidence. Secondly, considering costs later, after an assessment of clinical effectiveness, tends to focus decisions on a drug's price rather than on its overall value.

It seems perverse, however, to list the absence of cost-effectiveness as a major limitation of the usefulness of technology assessments when key decision-makers, such as Medicare

and private health plans, have neither the tools nor the stomach to be able to apply cost-effectiveness analysis explicitly. Some of these decision makers are quite frank that they have no interest in cost-effectiveness at all; clinical effectiveness is hard enough to determine in their view. Certainly, if insurers in the United States were ready and able to apply cost-effectiveness or other measures of value to benefit, coverage, reimbursement, or physician compensation systems, then technology assessment programs would meet their demand. The limitation represented by the absence of cost-effectiveness, therefore, is one for which the answer will require innovations across both producers and consumers of technology assessment production. Technology assessment programs will need to learn how to provide cost-effectiveness information with an objectivity, legitimacy, and transparency that meets the needs of decision-makers. Assessment programs will also need to develop new ways to structure the presentation of cost-effectiveness information so that it can be used as a managerial tool by decision-makers. But all this will only help decision-makers if they take matching strides to develop new methods for integrating information about cost-effectiveness into all of the methods they currently use to manage the value of health care.

Incomplete Integration

Thus the full promise of technology assessment to improve health care value can only be achieved if new approaches can be developed to integrate assessments thoroughly into areas where they have had little influence to date: the structure and function of insurance benefits, patient and clinician decision-support, medical management, pricing, and physician compensation. These are the main levers through which quality and efficiency are managed in the current health care system, and technology assessments currently play a very limited role in any of them. Medicare and private health plans apply health technology assessments almost exclusively to support just two functions: the initial coverage decision for a newly introduced technology, and the creation of tiered formularies. The assessments of TEC and DERP are structured to meet these needs alone. If assessments can be tailored to give decision-makers new ways to use information on the clinical effectiveness and on the comparative value of technologies, then there may be many additional ways to integrate them into multiple value-oriented strategies.

Integration is also incomplete between technology assessment efforts and the national processes within AHRQ, NIH, and other research agencies to prioritize and conduct new research. Technology assessment often highlights the specific questions for which the existing evidence is weakest, suggesting critical opportunities for further research to affect medical practice and policy. A new technique called “value of information analysis” even allows for technology assessments to model whether the expected impact of information from new research would represent a good investment of limited research funds. But there has been no major effort to channel the results of technology assessments performed by AHRQ or other major entities into a formal process of research prioritization. Technology assessment reports often include a section on suggested further research, but more often than not these sections are crafted by systematic reviewers without participation or consultation from experienced researchers.

As a result the recommendations for further research are often too numerous, overly vague, and ultimately fail to have a strong impact on the nation's research agenda.

Improved integration between technology assessment and efforts to prioritize and shape new research would also offer the advantage of speeding the incorporation of new research evidence into updated assessments, translating research results quickly into actionable information. If the two key functions of comparative effectiveness --- evidence development and technology assessment --- could be formally linked and better integrated, any benefits of increased funding and coordination would support and reinforce each component, strengthening their single and joint abilities to improve the value of health care.

IV. Centralized Technology Assessment Programs: The Lessons of NICE

In order to frame ways forward to address the four broad limitations of health technology assessment in the United States, it will prove instructive to examine what can be learned from the experience with technology assessment in other countries. The singular feature that distinguishes the international experience from that of the United States is the steady progression toward greater centralization and coordination. Australia, Canada, Germany, the Netherlands, Sweden, Norway, Denmark, France. These are just some of the countries that have developed significant national efforts in technology assessment. Australia has pride of precedence, with its Pharmaceutical Benefits Advisory Committee (PBAC) formed in the 1950s to assess drugs for inclusion on the national formulary. Canada and the other nations also have robust and in some ways unique systems, but this paper will concentrate on the lessons from the example of the United Kingdom. The National Institute for Health and Clinical Excellence (NICE) was established in 1999 to provide health professionals in England and Wales with advice on securing the highest attainable standards of care for patients in the National Health Service. Since its inception NICE has gained worldwide prominence for the rigor and independence of its technology assessment process. It also is viewed as establishing a system of unique transparency and openness to stakeholder involvement. NICE provides a striking contrast to the technology assessment in the United States through its reliance on cost effectiveness as the fundamental basis of comparison between new technologies and their alternatives. For these reasons, and for its tested political durability, NICE has become the most influential technology assessment initiative in the world, and it offers several intriguing and important lessons for efforts to improve technology assessment in the United States.

Structure, Function, and Methods

NICE has a broad mandate to set standards for the use of new technologies and procedures within the NHS and to produce guidelines for clinical, and now public, health. Details about the processes used in the development of these various forms of NICE guidance can be found elsewhere, and they have many common features, but here emphasis will be given to the operations of the center for technology appraisal, in many

ways the flagship initiative at NICE. NICE methodology makes an important distinction between technology assessment and technology appraisal. Assessment puts the evidence together; appraisal judges it in the context of a decision to fund or not fund the technology across the NHS.

NICE commissions its technology assessments from one of several independent academic groups in the UK (much as AHRQ does from its network of EPCs). Technology assessment follows a standard and well-described methodology that includes a full systematic review of the topic and a rigorous approach to economic modeling of cost effectiveness. NICE provides a sterling example of how an assessment program can combine usefully the methodologies of systematic review and decision modeling; more will be said below about the Institute's use of cost-effectiveness.

The appraisal phase sees the technology assessment delivered back from the academic unit to NICE, but only so that the evidence can be considered there by an independent advisory committee, chaired by NICE staff but whose members are drawn from clinicians, professional groups, researchers, and individuals with experience in patient advocacy. Based on its deliberations, the advisory committee makes its recommendation on funding to the NHS. Although health ministers in the government have reserve powers to advise the NHS to ignore NICE guidance, they have never done so.

Stakeholders (including relevant professional and patient organizations as well as manufacturers) are involved at all stages, from the preliminary scoping exercise that establishes the boundaries and comparator technologies for the appraisal, through the assessment phase when they have full access to the supporting systematic reviews; and on to the appraisal phase, when they are encouraged to comment on draft forms of guidance. There is also a formal appeals mechanism for stakeholders which, until just this past year, had proven robust enough to ward off formal legal challenges. Finally, the Institute attempts to ensure that its processes are as transparent as possible: NICE's work programs are publicized well in advance; and the data from which its conclusions are drawn are in the public domain with the exception of the details of studies that manufacturers insist remain "commercial-in-confidence."

Economic Evaluation

Of all the elements of NICE guidance the most distinctive, from a US perspective, is its use of economic evaluation to help judge the value of technologies that provide additional benefit but at an increased cost. The key measure used by NICE to assess the marginal value of a technology, for different patient groups, is the additional cost per quality adjusted life year (QALY) gained. If appropriate data on quality of life are unavailable, cost-effectiveness is estimated using alternatives such as the cost per life year gained.

Whether to recommend the use of a technology for certain patients and indications depends, in part, on the point at which the incremental cost per QALY is judged to no

longer be “cost-effective.” Recognizing this central feature of economic evaluation, NICE has carefully described its approach and expects its advisory bodies to use estimates of cost effectiveness to inform, but not determine, their decisions. In other words, NICE does not have a specific cost per QALY threshold above which a technology is rejected. Although research is continuing in this area, there is currently no empirical basis for assigning a specific cost per QALY threshold within the NHS; and, even if there were an empirical guide, an explicit threshold would suggest that health utility as measured by QALYs has absolute priority over other objectives, including various forms of equity. Nevertheless, NICE has arrived operationally at a band of approximately \$30,600 – \$45,900 per QALY (based on purchasing power parity of US\$1 = £0.65) as the threshold above which it would be increasingly likely to reject a technology on grounds of cost-ineffectiveness. For example, the Institute has approved the use of etanercept and infliximab, both with incremental cost effectiveness ratios of \$47,430 per QALY, in the treatment of rheumatoid arthritis; but it has rejected anakinra with an incremental ratio of \$102,510 per QALY.

Adopting this range of \$30,600 – \$45,900 per QALY as a benchmark for cost effectiveness maintains consistency across the many different types of health care technologies that NICE appraises and, at the same time, gives NICE’s advisory bodies latitude to consider the degree of uncertainty surrounding the estimate, the particular features of the condition, the innovative nature of the technology and, where appropriate, the wider societal costs and benefits.

It is important to note that NICE does not take the budget impact of a new technology into account. For example, although a new drug might have a favorable cost-effectiveness ratio of \$20,000/QALY, the overall impact might be quite significant for the NHS budget if large numbers of patients were to be eligible for treatment. The recent approval by NICE of Herceptin has created just such a bind for NHS budgets. Although the drug itself was found to be cost-effective, its relative high cost and the increasing number of patients eligible to take it have created a situation in which it has been estimated that 25% of the entire budget for cancer care in England could be spent on Herceptin. Nevertheless, NICE’s methods continue to ignore issues of affordability; the government remains accountable for the overall NHS budget and therefore has the responsibility to judge a particular intervention unaffordable for the NHS even though NICE might have judged it cost-effective.

Lessons of NICE for technology assessment in the United States

As a centralized technology assessment program, operating with meticulous attention to the quality and rigor of its products; as an organization whose methods project high degrees of objectivity, stakeholder engagement, and transparency; and, as a vanguard model of an explicit and actionable approach for considering cost-effectiveness of new technologies, NICE serves as an exemplar to which the United States can turn for lessons that may help address the deficiencies in its technology assessment programs. That does

not imply that NICE is flawless or that it provides a readily suitable model for technology assessment in the United States; but the experience of NICE has answered several important concerns about the potential application of technology assessment to efforts to improve health; and the obvious contrast between NICE's structure and function and the diffuse approach to technology assessment in the US provides a clear backdrop from which several specific lessons stand out.

First, centralization has worked in the UK. NICE's role has had a strong influence on establishing norms and has set a new standard of excellence in the methods of technology assessment. The format of NICE evidence reviews and guidance documents has created in many ways a common language for reviewers and stakeholders, adding to the transparency of the assessment process. NICE has also demonstrated that it is possible to function as a centralized program while maintaining a refreshing degree of political independence. It is structured as a "special health authority" within the NHS, which makes it accountable to the NHS, and, ultimately, to Parliament, but which allows it to operate with day-to-day freedom from political pressure. This independence has been tested but has proven resilient. Structural analogies for a similar kind of protected and accountable entity within the United States government are difficult to find, but NICE's success speaks to the possibility that difficult and contentious questions of evidence can be addressed within a federal political structure. Specific organizational models of a centralized technology assessment and comparative effectiveness entity in the United States will be discussed in the final section of this paper.

NICE's position and the timing of its assessments also suggest answers to the important question of whether technology assessment is feasible at or near the time of entry to market for new technologies. Critics of NICE have long maintained that the best time to assess the effectiveness and cost-effectiveness of technologies is not at the time they are first introduced but after they have been used in broader populations for at least several years. Although this argument is not unreasonable on theoretical grounds, NICE has shown that, for some technologies, a program of assessment near to the time of introduction into the marketplace can be done with enough rigor to satisfy decision-makers and help them manage value at the critical juncture of first coverage/funding. To be clear, NICE's technology appraisals have focused almost entirely on new drugs, which must have a portfolio of relatively robust data on safety and efficacy in order to gain regulatory approval at the UK equivalent of the FDA. These early data often provide enough information to allow NICE to use its combination of systematic review and decision analytic modeling to create a rigorous platform for assessment. But NICE does not attempt the same kind of assessment on all new technologies. NICE applies a different assessment process for new interventional procedures, which are introduced into medical practice without the kind of formal licensing procedure required of new drugs, and therefore come arrive at assessment's door step with far less and lower quality data to evaluate. Faced with the dearth of data, the NICE interventional procedure program assesses only safety and effectiveness; modeling to estimate cost-effectiveness is not attempted. Moreover, diagnostic devices are not formally assessed at all within the NICE set of programs. Thus the example of NICE demonstrates that assessments of new drugs shortly after introduction can prove robust in an assessment program that has a strong

base of analytic modeling, but NICE's experience cannot be used to support early assessment of all health care technologies, and casts some doubt on whether the cost-effectiveness of new procedures and devices should be included within an assessment program.

The specific technical and procedural methods of technology assessment at NICE provide several important lessons that could inform the creation of an enhanced, centralized technology assessment function in the US. NICE staff credit their own early survival and later success to several core features of their methods: 1) the insistence on rigor and quality of all products; 2) the clarity and transparency of their procedures; and 3) their willingness to engage with all relevant stakeholders. These elements have been described above but it is worth emphasizing again the importance of clarity in the assessment procedure. NICE provides schematics and clear timelines for all of its assessments, allowing stakeholders to always know what the "rules of the game" are, and giving them ample opportunity to prepare their evidentiary and other contributions to the process. Prior to NICE, manufacturers had been faced with dealing with decision-makers spread throughout the various districts of the NHS, each of whom might use different methods and different procedures for making a funding decision. With NICE there was a trade-off: a much bigger risk should the funding decision be "no," but in return manufacturers gained the advantage of being able to work with a single, sophisticated, transparent, partner.

The most difficult aspect of NICE to extrapolate to the US setting is the thorough grounding of its technology assessment process in cost-effectiveness. The resistance to cost-effectiveness in the United States, particularly within the Medicare program, is legendary. However, proponents of cost-effectiveness find in the experience of NICE the proof that cost-effectiveness, despite its technical difficulties and many inherent problems, can assume the legitimacy necessary to provide a foundation for a national technology assessment program. After all, if the goal of technology assessment is to aid decision-makers in obtaining better value, then costs have to be considered in some fashion. Formal use of cost-effectiveness allows NICE to be so explicit about the role of costs that it can avoid the suspicion that hovers about AHRQ, TEC, DERP, and all private health plans – that costs are being considered covertly in the judgment of the evidence of clinical effectiveness. By contrast, when NICE sets out to assess the comparative value of a new technology it is doing so in a comprehensive and cohesive manner. NICE still struggles to make the cost-effectiveness models it uses transparent enough and reliable enough to stand up to scrutiny, but that is where the overall procedural transparency and independence of the Institute play such a key facilitating role in its success.

It is useful to identify the key aspect of NICE that specifically would not be reasonable to consider replicating in the United States. NICE makes decisions for the entire NHS; and its decisions have to take into consideration all the different types of treatments it appraises year over year. Thus there is perfect justification for the way that NICE not only generates information on the cost-effectiveness of new technologies, but applies a broad but essentially stable cost-effectiveness threshold in rendering its decisions. The

US health care system is pluralistic, and there is therefore no appetite for a centralized, unitary decision-maker. A centralized, objective, and authoritative source of information --- yes; but in the US that information would more plausibly be made available to a pluralistic set of decision makers who would use the information to provide choice. How the elements of technology assessment could be assembled to provide the basis for choice will be a key theme of the next section.

V. Redesigning Technology Assessment in the United States

Having now surveyed the spectrum of major US technology assessment efforts and mined the experience of the foremost international HTA organization for relevant lessons, we must revisit the list of broad limitations that impede the ability of technology assessment to improve the value of health care in the United States. This list includes 1) Poor Coordination; 2) Weak Legitimacy; 3) Limited Usefulness; and 4) Incomplete Integration. The recommendations presented below will present ideas for changes to technology assessment in the United States that will address each of these limitations. The recommendations will describe a more coherent and cohesive approach to technology assessment than exists today. Greater consistency and rigor in methods will characterize all programs, along with a more explicit emphasis on transparency and stakeholder involvement. Importantly, cost-effectiveness and other measures of value will be seamlessly integrated into technology assessment platforms, providing decision-makers with information on comparative value in a more explicit and actionable form. For their part, decision-makers will take new strides to use the information from technology assessment as a core element with which to design innovative health care benefits, pricing strategies, physician compensation programs, and patient-clinician decision-aids. The keystone of this new system, presented as the final recommendation, will be a newly created federal entity that will integrate technology assessment and comparative effectiveness research; a “trusted resource,” whose rigor, objectivity, transparency, collaborative philosophy, and usefulness to decision-makers will provide durable leadership and harness the full power of technology assessment to help improve health care value.

Recommendation 1

Further policy development should seek to improve the rigor and consistency with which the strength of evidence is assessed and judged to be adequate for key health care decisions. Standard nomenclature and definitions for categories of the strength of evidence should be developed and linked conceptually to a clear framework for integration into coverage and reimbursement policy.

As noted earlier, the legitimacy of technology assessment hinges on perceptions of objectivity, the opportunities for stakeholder involvement, and the transparency of procedures and of decision-making. This first recommendation offers opportunities to improve all three of these elements. Active research and policy development needs to

establish clear standards for how technology assessment efforts should evaluate the strength of evidence. A common approach and terminology would not eliminate differences of opinion regarding the strength of evidence on the effectiveness or value of a new technology, nor would all coverage and pricing decisions achieve perfect harmony. But even the seemingly simple step of adopting common terms to categorize the comparative clinical effectiveness of new technologies would have a profound effect. Direct comparison of technology assessments across sources would be made more feasible. Deliberation around limitations in evidence would become much more transparent. And, if a common terminology can be made the product of a common framework through which assessment groups would weigh the evidence, this would lay the groundwork needed to improve the transparency and consistency of medical policy decisions.

A consistent and rigorous framework, entailing a common set of terms and definitions, would therefore help make technology assessment far more transparent than it now is. This is part of NICE's legacy in the UK. Delineating a framework, or roadmap, for evidence-based decision-making in the US would be more difficult without a centralized, dominant technology assessment program, but even in a more pluralistic framework it might be possible to gain a critical number of key adherents and slowly build toward greater consensus. Manufacturers should definitely be involved, for they have much to contribute and much to gain; for them improved rigor and transparency translate into greater predictability of the assessments of their products. Progress in this direction would be instrumental in improving the legitimacy and transparency of technology assessment efforts.

There are several ongoing evidence-based medicine initiatives, both within the United States, and internationally, that may produce important advances. Among these is the Evidence-Based Medicine Roadmap Group, led by America's Health Insurance Plans (AHIP), with participation from AHRQ, CMS, Merck, Johnson and Johnson, Boston Scientific, several private health plans, employers, and patient advocacy groups. The work of this group to date has produced a draft "matrix" for defining the categories of strength of evidence on comparative effectiveness. This matrix is shown in Appendix A. Next steps will further define the boundaries between various thresholds in the standard of evidence and then describe a conceptual model for how decision-making bodies should integrate judgments of the strength of evidence into the broader process of making coverage and reimbursement decisions.

Recommendation 2

Technology assessment programs, including those in the private as well as public sectors, should develop procedures that meet high standards of transparency and stakeholder engagement. Explicit mechanisms for incorporating patients' and society's perspectives on effectiveness and value should be developed and clearly described.

In addition to further work on the methods of assessing and communicating the strength of evidence, technology assessment programs need to examine and adopt best practices such as those established by NICE and other top organizations to improve the transparency of their entire organizational assessment process. It is difficult to describe gradations of clarity and transparency in organizational processes, but it only takes a casual stroll through the NICE website to gain an understanding of how clearly an organization can describe all its assessment procedures, from topic selection to research and analysis standards; from timelines to templates for stakeholder input. The degree of transparency and stakeholder engagement that NICE has reached requires a commitment of significant attention and resources from a technology assessment organization, and it is clear that many smaller and private assessment programs will simply not have these available. But every major US program could identify priority areas where improvement in transparency and stakeholder engagement are both possible and feasible; such action would enhance the legitimacy of their programs and help raise standards across the diverse set of US programs.

Patients are obviously central stakeholders to any technology assessment process. Because patients are assumed to lack technical or clinical expertise while being “burdened” with highly personal experiences and motives, many technology assessment programs explicitly exclude them on the grounds of objectivity, or at best do not have well-developed programs for integrating patients and their families into the technology assessment process. Here again NICE has often been at the forefront of new methods. Patients participate in every stage of NICE technology appraisals and NICE has developed a Patient Involvement Unit to help train patients and other laypersons to participate excellent transparency procedures their transparency and expand the opportunities they provide for stakeholder engagement. NICE also has had for several years a Citizen’s Council which helps brings broader societal values directly into the assessment process. Technology assessment programs in the US should look for their own innovative ways to ensure that patients’ and society’s perspectives are fully represented in order to improve the validity and legitimacy of assessment.

Recommendation 3

The assessment of comparative value, which includes formal economic analysis, should become a core element of technology assessment in the United States. Economic analysis should be based on transparent decision analytic models that meet high standards for objectivity and technical quality. Manufacturers, patients, clinicians, and insurers must all be engaged in the determination of the proper focus and scope of the assessment of comparative value.

In order to be most useful in the effort to improve the value of health care, technology assessment must include within its mandate the evaluation of comparative value. Assessment of clinical effectiveness alone has been the dominant US version of technology assessment and it is important to reflect again on its strengths. Assessment of clinical effectiveness is can help insurers decide whether a new technology has enough evidence to justify consideration of coverage as “medically necessary.” Clinical effectiveness is also the paramount concern of insured patients and of clinicians. But costs do matter. Costs matter to some individual patients who bear financial responsibility for their care; costs in the aggregate matter to insured patients more generally who are watching their insurance coverage erode over time as costs outstrip employers’ abilities to subsidize employees’ insurance; and costs are of great concern to the employers and insurers who hold the lion’s share of financial responsibility for health care services. Thus, for the legitimacy and usefulness of technology assessments to advance, it will be necessary for technology assessments to grapple more explicitly not only with costs but with measures of the marginal value of technologies compared to established alternatives. Comparative value must become a key focus of technology assessment.

It is important that the concept of assessing comparative value not be construed as implying that value is monolithic and unalterable across the perspectives of different stakeholders. On the contrary, good technology assessments will highlight where the perspectives of patients, clinicians, insurers, employers, and society at large differ most on what counts as value, and on the weights assigned to each component of value. This is why it is so important that the focus and scope of all technology assessments be determined through an open process that includes all these major stakeholders. NICE has had the greatest experience with the “scoping” of technology assessments to clarify the distinctions of various perspectives on value, and they have learned that the ultimate legitimacy and usefulness of assessments is often strongly determined by the thoroughness of this scoping phase. There will be a primary, or dominant, perspective on value that will be selected as the norm for a technology assessment program. For NICE the dominant perspective is that of the health service. In the US, the dominant value perspective is often assigned to whoever commissions the technology assessment, in many cases an insurer. It is likely that in a national program the primary perspective will be that of society as a whole, but it will be important for assessment reports to report other perspectives as well.

One question that often arises when considering economic analysis as a component of technology assessment is whether there will be a uniform metric of comparative value, the obvious candidate being the cost/QALY used by NICE and long dominant among health economists. There are significant theoretical advantages to having cost/QALY as a single metric. A single metric allows technology assessment to compare value across types of technologies and over time. Standardization aids in the transparency and legitimacy of the process, since decision makers using technology assessments would need to justify their decisions based on the same metric of value. In addition, if the US were to adopt the cost/QALY as a core element of a new national technology assessment initiative, the US would join a growing international consensus and further strengthen the underlying methodological basis for economic analysis. This consensus would help send a clear signal to manufacturers that data on patient utilities as well as clinical outcomes would be included in the assessments of the most important medical market in the world, spurring them to gather such data as part of their development of all new technologies.

But it may be that in the US context such a full-fledged commitment to the cost/QALY would be unwise and unnecessary, at least the outset. In part this is because no technology assessment program, not even the new federal initiative envisioned by the final recommendation of this paper, will be making coverage or reimbursement decisions. Assessments of comparative value might, therefore, be more naturally judged to lie in other metrics, such as cost per life year gained, or cost per consequence (e.g. cost per additional stroke prevented). This last version of a comparative value metric, cost per consequence, is more intuitive than cost/QALY to clinicians and patients. Particularly in situations in which the data on utilities is unreliable it may be wise to opt away from cost/QALY as the dominant metric of comparative value for a particular technology. This paper recommends, therefore, that cost/QALY be estimated for every technology for which the data make the estimation reasonable, but that the scoping process, in which all stakeholders participate, establish at the initiation of the assessment whether the cost/QALY or a different metric of comparative value will be the primary metric reported.

Recommendation 4

The usefulness and impact of technology assessments in the health care system should be advanced by innovations in three areas: 1) collaboration with manufacturers to improve timeliness; 2) inclusion of decision analytic modeling as a complementary basis for technology assessment; and 3) the adoption of new models for framing and formatting information to help decision-makers more clearly understand the clinical effectiveness and comparative value of new technologies.

It takes time to do a thorough review of the medical evidence and prepare a formal technology assessment. Creation of decision analytic models, and inclusion of economic analysis, can extend the timeline for assessment to a year or more in many academic settings. For many key decision makers in the US health care system this is just too long, for they must make coverage and reimbursement decisions relatively soon after the

introduction of a new technology into practice, and often those initial decisions are difficult or impossible to modify later on. The problem of timeliness also is critical for manufacturers who potentially lose access to customers and revenue every day that a technology is “under review.”

Greater coordination and support for the infrastructure of technology assessment is one avenue to improve timeliness. Another, less explored option is greater collaboration with manufacturers. Concerns about the influence of manufacturers on the assessment process have made most researchers and administrators of technology assessment programs leery of direct collaboration. There is empirical data to support suspicions that economic models designed by manufacturers produce more favorable estimates of cost-effectiveness than do models created de novo by independent academic units. Nevertheless, manufacturers often have superior knowledge of their own product and its comparators. Manufacturers also have strategic and competitive interests that often lead them to develop sophisticated economic models of the clinical effectiveness and comparative value of their product. Further work needs to be done to develop transparent and rigorous collaborative relationships between assessment programs and manufacturers so that the information and expertise of manufacturers can be used to create more expeditious and efficient assessment processes. This recommendation extends far beyond the idea of a “user fee” such as that imposed by Congress on manufacturers seeking regulatory approval of new drugs at FDA. NICE has recently had to respond to concerns about its timeliness by creating a new fast track for assessments of certain single technologies. This “STA” process takes approximately one half the time of former NICE appraisals and can do so primarily based on the requirement for manufactures to submit a working economic model with full justification at the beginning of the appraisal process. NICE is in the early stages of learning how well its manufacturer submission template and other elements of the new STA process serve to maintain the rigor and legitimacy of their appraisal. But NICE is a harbinger of the fact that traditional academic timelines for the performance of technology assessment have appeared unhelpful to many decision makers are unlikely to meet the needs of any future enhanced national assessment program in the US. Greater collaboration with industry to speed the overall process will help make technology assessment a much more practical tool to improve value.

The usefulness of technology assessments should also be enhanced by adopting new ways of framing and formatting information for decision-makers. For patients and clinicians this can include relatively simple calculations from the available data of measures such as “number needed to treat” and “number needed to harm.” Quantitative and qualitative estimates of net health benefit and the certainty with which certain benefits and harms are known are also very important for many decision makers. The EBM Roadmap initiative mentioned earlier is seeking to create a link between types of evidence and readily recognizable categories, such as “superior” “incremental” and “comparable” that will be more accessible and therefore more useful.

The most innovation is needed, however, in ways to communicate comparative value to decision makers in a format that retains rigor while yielding to interpretation by those without a PhD in health economics. One model for such an approach is being developed

and tested in the US by the Institute for Clinical and Economic Review (ICER). The methodology of ICER will be described here as a concrete example of a model for improving the usefulness of technology assessments by formatting the information in a way that can offer multiple opportunities for inclusion in efforts to improve health care value. ICER is an academic initiative launched in 2006 and based at Harvard Medical School. ICER has developed a method which formally links assessments of clinical effectiveness to those of comparative value in a multi-dimensional rating system, called the Integrated Value Rating (see Figure below).

Integrated Value Rating (IVR)

Comparative Clinical Effectiveness

Superior	A	Ac	Ab	Aa
Incremental	B	Bc	Bb	Ba
Comparable	C	Cc	Cb	Ca
Promising	P	Pc	Pb	Pa
Uncertain	U	U_	U_	U_
Comparative Value		c Poor	b Reasonable/ Comparable	a Superior

The first element of the IVR (shown in the Figure as the y-axis) emerges from a systematic review of the medical literature to determine the comparative clinical effectiveness of a therapeutic agent. The clinical effectiveness review summarizes key information about the benefits, risks, and clinical applications of the technology being assessed. The technology will be given a rating for clinical effectiveness as being “Superior, Incremental, Comparable, Promising, or Uncertain.” Each of these categories represent a combined judgment of the magnitude of the net health benefit provided by the technology and the relative strength of the evidence supporting that benefit.

The second piece of the IVR (shown on the x-axis) is an analysis of the comparative value of the technology. ICER uses economic models to provide an overall assessment of the incremental costs and benefits of the new agent compared to other reasonable alternative treatments. As discussed earlier, ICER will calculate cost/QALY where feasible but may be directed by its scoping groups to focus more closely on cost/consequences or some other metric of value. It is important to note that whatever formal cost-effectiveness is done is an input into ICER’s approach to rating comparative value, but is not the sole factor. Consideration is also given to how the costs and benefits

compare to treatments with similar benefits for other diseases. In addition, the comparative value rating acknowledges that certain therapeutics, including some medical devices and surgical procedures, have an evolutionary cycle likely to result in near-term cost reductions. Ultimately, in order to be of greatest use to decision-makers, ICER rates the comparative value of a new therapeutic within one of three broad categories, Superior, Reasonable/Comparable, and Poor.

Some stakeholders, including some patients and clinicians, may be primarily interested in the piece of the assessment that looks at comparative clinical effectiveness, with less or no attention to comparative value. But the overarching goal of the ICER method is to give those stakeholders who wish to know and use information on comparative value an integrated assessment that gives them many options for application. Putting the ratings for clinical effectiveness and comparative value together gives a complete Integrated Value Rating for a technology. As shown in the Figure, the 5x3 grid of possible IVR combinations of ratings for clinical effectiveness and comparative value provides an evidence-based “menu” which can support serves multiple purposes. The specific possible uses of this kind of technology assessment format will be discussed in the next recommendation.

Recommendation 5

Public and private insurers should work with employers, clinicians, regulators, and patients to develop innovative ways of incorporating technology assessments into the full menu of programs used to improve the value of health care, including:

- a. benefit design*
- b. coverage (e.g. CED)*
- c. pricing/reimbursement and medical management*
- d. physician compensation*
- e. patient-clinician decision-support systems*

One of the broad limitations of technology assessment in the United States described earlier was the incomplete integration of technology assessment into many of the methods used to manage and improve the value of health care. Now that we have developed the concept of technology assessment that can provide a joint evaluation of both clinical effectiveness and comparative value, it is important to discuss how this information could be used.

One important application of the ICER matrix of technology assessment ratings is in the area of benefit design. Insurers have used considerations of cost and effectiveness to design tiered drug formularies, but have otherwise not had a practical model for tiering the benefits related to procedures, devices, and many specialty services. Instead, many benefit designs have relied on a “blunt” deductible that applies equally to all health care services, no matter what the effectiveness or value. ICER technology appraisals could be used to tier services, particularly in areas where there are several established options and

some reasonable degree of patient choice. Treatment choices for localized prostate cancer are one example of such a clinical area where benefit design could incorporate ICER ratings to tier treatment options.

ICER ratings could also be useful to support more innovative coverage decisions, including those linked to requirements to participate in ongoing clinical research (sometimes called “coverage with evidence development” or “coverage under protocol”). Both public and private insurers need a valid, objective way to decide which new technologies should receive this form of coverage, and it is possible that the integrated appraisal rating could prove the foundation on which this type of approach could grow with greater legitimacy.

Health plans and patient advocacy groups could use ICER appraisal ratings as the basis for reimbursement negotiation with manufacturers of a new technology. If a new technology is rated as only demonstrating “Incremental” comparative clinical effectiveness, then negotiation might undertake to bring the price down to a point where the comparative value rating would reach the “Reasonable/Comparable” threshold. Similarly, for technologies that rate as having “Superior” effectiveness and “Superior” comparative value, insurers could take special steps to lower any possible barriers to the rapid dissemination and adoption of this new technology.

Similarly, both clinician compensation and patient-clinician decision aids could be adapted to use integrated technology ratings as a method for improving the value of care. Pay-for-performance programs could be designed to use integrated ratings of new or old technologies as the basis for rewarding clinician groups that deliver a high proportion of “Superior/Superior” technologies to their patients. Health plans, specialty societies, and others could also use this type of technology assessment to help design decision support tools for patients and their clinicians.

In summary, redesigning technology assessment will yield the greatest impact on care only if the format of assessments is readily adaptable for use in many different elements of the health care system. A rating system such as the one developed by ICER, in which clinical effectiveness and comparative value are both readily identifiable and yet also integrated, offers the mixture of rigor, consistency, accessibility, and flexibility of application that may help technology assessment assume a much more important role.

Recommendation 6

Congress should act to create a new federal entity to coordinate and conduct technology assessment as a public good. This new entity should be responsible for (1) comparing new and existing procedures, therapies, medical devices and other technologies for effectiveness; (2) assessing alternative uses of standard interventions; and (3) conducting and linking reviews of clinical-effectiveness and cost-effectiveness to empower clinicians and consumers to make more informed decisions regarding the value of health care interventions.

This new federal entity entrusted with leading technology assessment should have the following functions, some of which it will share or delegate to governmental or private bodies:

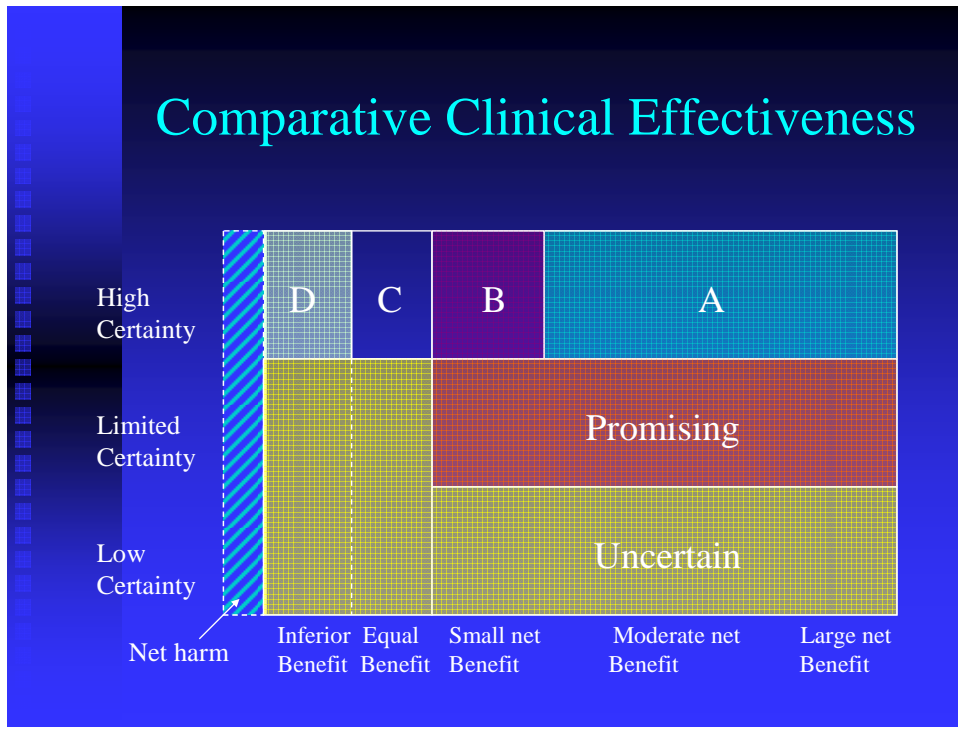
- 1. Prioritize technologies for evaluation*
- 2. Systematically review existing evidence on clinical effectiveness*
- 3. Fund and/or conduct studies of comparative clinical effectiveness*
- 4. Evaluate cost-effectiveness or other value measures*
- 5. Set methodological standards*
- 6. Disseminate results to clinicians and patients*

There are several reasonable options for placement and funding for this new entity. The ultimate decision should proceed with great attention to the perspectives of all stakeholders, since the future success of the entity is rooted in the base of support and engagement it can build upon from its inception.

All of the other recommendations made in this paper will achieve their greatest impact only if there is federal support for a leading national entity for technology assessment. A centralized leader would immediately lend cohesiveness to the development of more standardized methods and to the prioritization of technologies for assessment. A federal entity could enlarge the scope of topics of technology assessment to include more existing and potentially obsolete technologies. The critical concern regarding objectivity, the concern that undermines much of the potential of technology assessment in this country, could be greatly reduced in the work done by this entity. It would create information, useful and actionable; it would be able to sponsor collaborative research on comparative value as a core component of its activities; it would have the authority and prestige to work effectively with manufacturers; and its leadership would give increased impetus to the use of technology assessments in benefit design, coverage, pricing, compensation, and decision supports.

The design of the new entity should be customized to fit the U.S. health care system. It should be accountable to Congress yet structured so that it has a political insulation similar to that which has allowed NICE to survive. The new organization should be independent, be funded by both the private and public sectors, and provide the highest quality assessment of evidence on clinical effectiveness and comparative value, but without the power to link these assessments to coverage or pricing decisions. The information then would be made available to practitioners, consumers, purchasers and health plans to support better treatment decisions.

Appendix A. The EBM Roadmap Group Draft Evidence Matrix



Our approach to assessing the comparative clinical effectiveness of a new technology is shown diagrammatically in Figure 1. We assume that for most medical policy decisions the clinical effectiveness of the technology is being compared to the most appropriate specific “comparator” already in use, but our model is also suitable for assessments of the evidence comparing a technology against placebo or “supportive care.” The comparative clinical effectiveness (CCE) model graphs the rating of a new technology along two axes. The x-axis is net health benefit, and the y-axis is the level of certainty provided by the body of evidence. Net health benefit begins in a zone in which the body of evidence suggests that the technology is more harmful than beneficial. As the relative net health benefit increases past zero net health benefit, it enters a zone in which it is judged to be net positive but still less positive overall than the net health benefit provided by the comparator. As net health benefit increases in the graph, the assessment indicates that the technology provides equal benefit to its comparator, then a small advantage in net health benefit, continuing on the axis toward a moderate-to-large advantage. The reason to have these categories is made clear when viewed in conjunction with the y-axis of the CCE model, which presents the level of certainty separated into three broad categories: high, limited, and low. As can be seen in the figure, when these three categories of certainty are mapped upon the categories of net health benefit, a matrix is revealed that can be used to define a specific taxonomy of comparative clinical effectiveness. A technology whose evidence base provides high certainty of a moderate-to-high net health benefit is rated to have “superior” comparative clinical effectiveness. Similarly, limited certainty of either an incremental or moderate-to-high net health benefit is called a technology with

“promising” comparative clinical effectiveness. The matrix design produces seven different ratings of comparative clinical effectiveness to guide deliberations and ultimate medical policy decisions.

The ultimate usefulness of any rating scheme such as our matrix lies in the degree of clarity and consistency it can bring to assessments and, ultimately, to the subsequent appraisal phase. Thus it is very important for us to clarify the boundaries that separate the categories in the matrix.

- What, for example, are the distinctions through which an assessment effort can reliably determine whether the level of certainty provided by a body of evidence is high, limited, or low?
- Similarly, what is meant by a “small” net health benefit as opposed to a moderate-to-high benefit?

Clarifying these boundaries, and giving examples that can demonstrate how these boundaries can be reliably determined in practice, will be a central part of the future work on the matrix as we move forward. We should be clear, however, that there are important limitations to this approach. Descriptions of the boundaries of comparative clinical effectiveness should be detailed and complete enough to help assessment groups improve the reliability and consistency of their assessments of evidence, but we recognize that an important degree of scientific judgment will remain in any classification scheme. Any attempt to define the matrix boundaries with overly specific numbers or types of studies, with untested quantitative measures of evidentiary consistency, will only create inflexibility that will inadequately serve decision makers facing the great variety of evidence bases and the types of questions that are being addressed. We also believe that the boundaries between the categories of our matrix will differ specifically depending on what kind of health care intervention is being evaluated. Future work of the EBM Roadmap group will explore this issue and seek consensus on how best to identify the boundaries in our matrix across different types of technologies, including therapeutics, procedures, devices, and diagnostics.

Appendix B. ICER Evidence Review Process**ICER™ Evidence Review Process****Topic Selection**

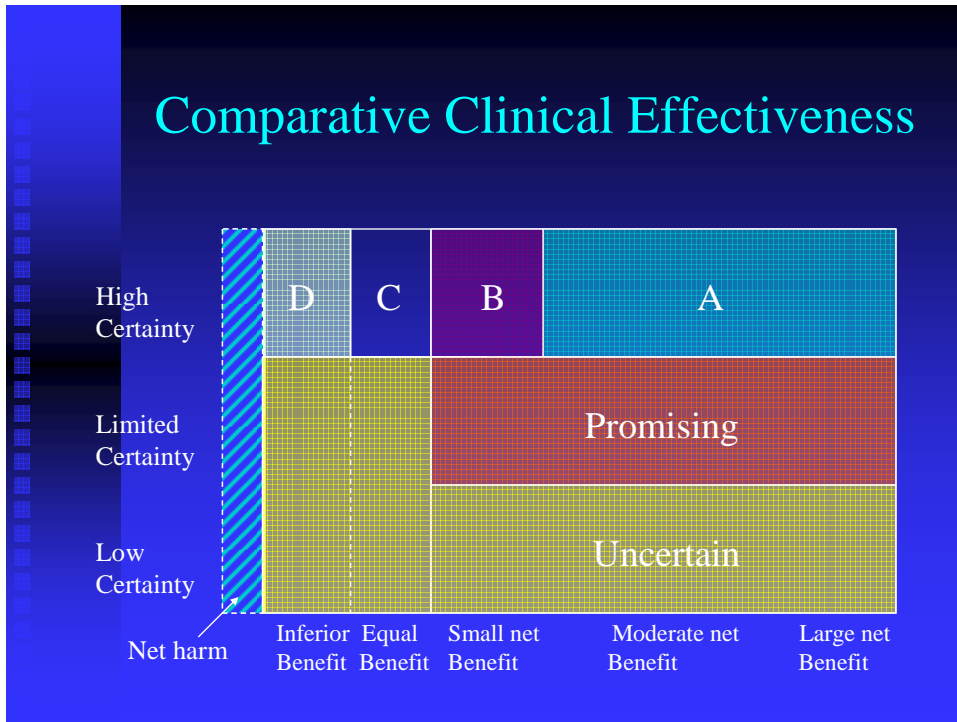
- Horizon scanning list prioritized by topic selection committee composed of representatives from all stakeholder groups: purchasers, payers, providers, patients.

Scoping

- Scoping committee formed for each topic comprised of clinician experts, patient representatives, ICER staff, payer representative, and industry representative(s)
- Scoping workshop held to specify:
 - a. “boundary” and benchmarks for risks, benefits, and costs to be used
 - b. comparator(s)
 - c. key patient subgroups
 - d. outcome measures
 - e. time horizon

Evidence Review and Integrated Value Rating (IVR™)

1. *Introduction*
 - a. History of technology development/cycle
 - b. FDA or other regulatory status
 - c. Clinical use (i.e. sequencing with other alternatives, etc.)
 - d. Prior coverage or other relevant history
2. *Evidentiary and ethical contextual considerations*
 - a. Severity of condition
 - b. Last chance or only effective therapy for severe condition
 - c. Vulnerable population or previous inequity in access/treatment
 - d. Unique patient/clinician perspective on added value
 - e. Cost-effectiveness estimates for “comparable” treatments or services related to “comparable” illnesses
3. *Comparative Clinical Effectiveness*
 - a. For each key patient subgroup:
 - i. type and magnitude of risk, benefits, and net benefit (including absolute risk reduction, NNT, etc.)
 - ii. Summary clinical effectiveness rating for net health benefit



Comparative Clinical Effectiveness Rating	Level of Certainty	Magnitude of Net Comparative Health Benefit
A “Superior”	High	Moderate-Large
B “Incremental”	High	Small
C “Comparable”	High	Equal
D “Inferior”	High	Inferior
P “Promising”	Limited	Small-Large
U “Uncertain”	Limited-Low	Equal-Large

4. *Comparative value*

- a. For each key patient subgroup:
 - i. Incremental cost per key outcomes, e.g. hospitalization avoided, death
 - ii. Incremental cost-effectiveness ratio: cost/QALY or cost/LYG

Comparative Value Rating	
a “Superior”	Cost-saving with upper CI < \$50K and/or incremental cost per key outcomes significantly less than comparators
b “Reasonable/Comparable”	Point estimate <\$150K with upper CI < \$175K and/or incremental cost per key outcomes similar to comparators
c “Poor”	Point estimate >\$150K with lower CI > \$100K and/or incremental cost per key outcomes significantly higher than comparators

5. *Budget impact and delivery system issues*

6. *Summary: Draft Integrated Value Rating (IVR™)*
- a. Health benefits and risks for key patient subgroups. Summary of relative uncertainty on key outcomes.
 - b. Key parameters to which the comparative value is sensitive.
 - c. Summary proposed ICER two-part “integrated value rating” (IVR™) for each key clinical subgroup

Integrated Value Rating

Comparative Clinical Effectiveness				
Superior	A	Ac	Ab	Aa
Incremental	B	Bc	Bb	Ba
Comparable	C	Cc	Cb	Ca
Promising	D	Dc	Db	Da
Uncertain	U	U_	U_	U_
Comparative Value		c Poor	b Reasonable/ Comparable	a Superior

- d. Option for use in evidence-based cost-sharing, reimbursement, pay-for-performance, etc. For example:
- 1) Aa or Ab or Ba → low copay, high reimbursement, positive P4P incentives
 - 2) Ac → high copay or prior auth
 - 3) Bb, Bc → high copay, negative P4P incentives
 - 4) Pa, Pb → coverage only in research
 - 5) Pc, U → no coverage

Evidence Review Group Deliberation and Final IVR™

- Independent evidence review group of methodological and policy experts unconnected to specific stakeholder groups.
- Voting/consensus to determine final proposed IVR™ based on consideration and discussion of evidence review.