Creating a Center for Evidence-Based Medicine

A study conducted by staff from the American Institutes for Research for the Medicare Payment Advisory Commission

Marilyn Moon
Benjamin Smith
Sigrid Gustafson

The American Institutes for Research
1000 Thomas Jefferson Street
Washington, DC 20007

•

MedPAC
601 New Jersey Avenue, NW
Suite 9000
Washington, DC 20001
(202) 220-3700
Fax: (202) 220-3759
www.medpac.gov

•

The views expressed in this report are those of the authors. No endorsement by MedPAC is intended or should be inferred.

February 2008 • No. 08–1

Creating a Center for Evidence-Based Medicine*

Marilyn Moon
Benjamin Smith
Sigrid Gustafson

The American Institutes for Research

Produced for
The Medicare Payment Advisory Commission

July 2007

The authors wish to acknowledge the information provided by Phil Davies and Ruth Lopert and research assistance of Lauren Smeeding. Opinions expressed herein are those of the authors only and not of AIR nor its Board of Directors.

Improving health care through application of evidence-based analyses of healthcare goods and services has long been held up as an ideal "for the future." But how to apply the results of research, what standards of evidence are needed, and who should have what power to apply these results remains a controversial topic in the United States. Other countries, such as the United Kingdom, Australia, and Canada, are much further along in the application of evidence-based information for decision making. Thus far in the United States, much of the attention has focused on generating the tools of analysis, although a number of states, private managed care organizations and the Veterans Administration all have used some form of evidence-based decision making for establishing Medicaid prescription drug formularies. Medicare also makes decisions about adoption of new techniques based on evidence of effectiveness. What has been lacking is any universal effort to apply evidence more broadly, moving beyond effectiveness studies to identifying best practices and/or avoiding wasteful spending.

Despite considerable talk of research to develop practice guidelines, comparative effectiveness analyses and sometimes even cost effectiveness work, the United States remains a long way from seeing much in the way of practical applications. Resistance and sometimes hostility from key stakeholder groups (such as providers of care) resulted in attacks on the work done by the Agency for Healthcare Research and Quality (then known as the Agency for Health Care Policy and Research). Gray, Gusmano and Collins (2003) have suggested that the general disinterest of lawmakers in health services research, the identification of some of the work with the Clinton Administration's goals, and powerful enemies (particularly back surgeons after development of a controversial set of guidelines) nearly dealt a death blow to the agency in the late 1990s.

This paper considers what might be done as first steps toward a more comprehensive and systematic approach to using evidence for improving health care. In doing so, we consider some of the barriers and challenges facing such an effort, what a structure for an organization to promote evidence-based applications might look like, and lessons and cautions for moving forward. In particular, we consider the necessary components of an organization that might be established, drawing on lessons both from Britain's National Institute for Health and Clinical Excellence (NICE) and from a recent federal government effort, the National Registry of Effective Programs (NREP) developed by the Federal Center for Substance Abuse Prevention under the Substance Abuse and Mental Health Services Administration. The last section of the paper examines a range of practical considerations in creating a center for healthcare improvement. Finally, an appendix to the paper examines some of the measures used in comparative and cost-effectiveness studies; an additional barrier to establishing an organization to promote more analysis is the lack of firm consensus on the specific measures that should be used.

A CENTER FOR EVIDENCE-BASED MEDICINE

Conceptually, a national center for evidence-based medicine would be an independent entity that would systematically identify evidence-based practices, conduct rigorous independent reviews of evidence-based research using strict protocols guided by methodological criteria, and disseminate objective information. The review process would begin with published and unpublished evidence submitted by investigators responsible for the primary research. This model would include an evaluation of the quality of the research used to establish the evidence-based practice, by applying a standardized, consensus developed, set of methodological criteria. As a result, competing practices can be evaluated relative to clinical outcome and rigor of the original analyses conducted.

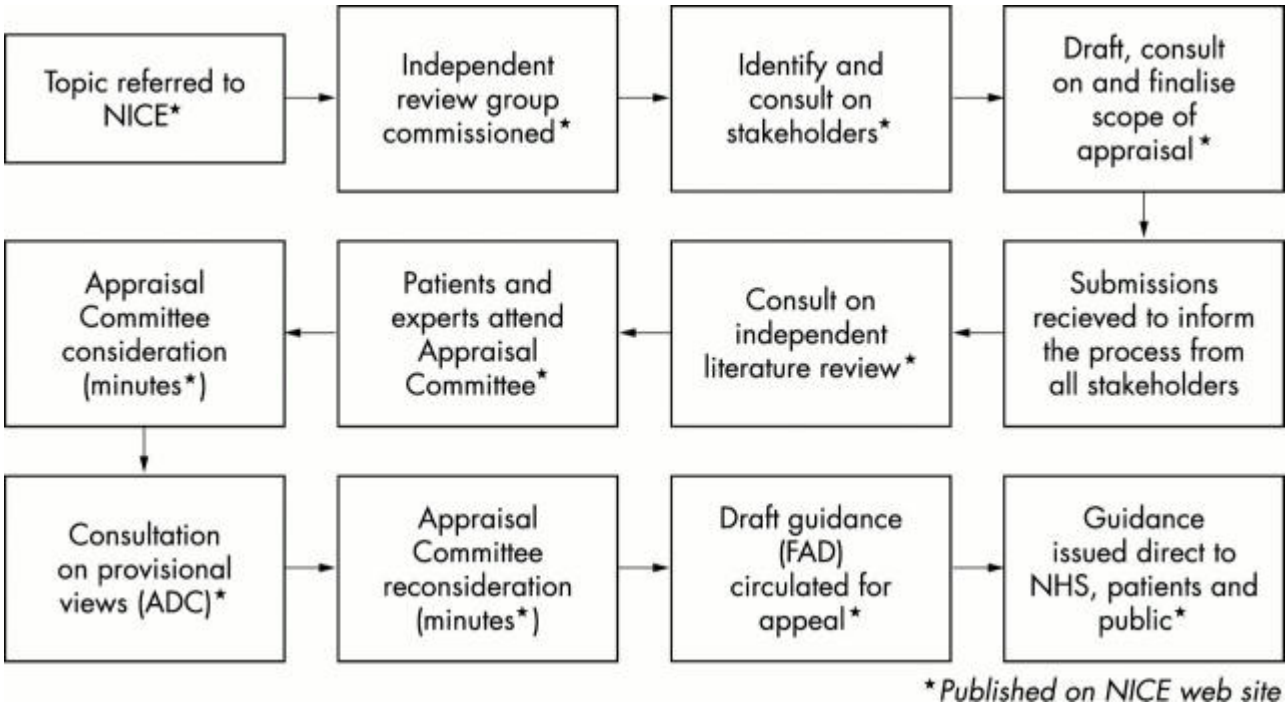
Certainly one model for the structure of an evidence-based center would be NICE—the National Institute for Health and Clinical Excellence—which was established by the British government in 1999. NICE is charged with providing "guidance" to health care professionals on the highest attainable standards of care for the National Health Service. (The term guidance was carefully chosen to reflect that NICE does not prescribe how its results will be applied. The National Health Service makes decisions about coverage.) It is fully funded by the British government but maintains substantial independence from political influence. Its mandate is broader than just offering the guidance; it seeks to reduce the variation in the quality and quantity of care delivered as well as performing clinical and cost effectiveness studies. These two characteristics of independence and a broader mandate to be viewed as the center for clinical excellence and standard setting put NICE in a good position to be effective. However, it is also criticized for a number of failings that may also be instructive, as described below.

The description of tasks for a Center for Evidence-Based Medicine (referred to below as the Center) is based roughly on the structure of the National Registry of Effective Programs (NREP) employed by the Federal Substance Abuse and Mental Health Services Administration (SAMHSA). Implicit in this model is a set of necessary activities to establish a process for review of clinical evidence, determination of efficacy and dissemination of highly ranked interventions (see the enclosed Chart 1). Since this represents an effort by the federal government to establish an organization to serve much the same purpose as that described here it is useful to begin with its basic elements. Its initial success stemmed from acknowledgement of the provider community about the value of reviews. After a change in focus, however, NREP's influence has declined, offering some sobering lessons about the importance of separating such

an organization from the general operation of government agencies. (See the box that describes the NREP in more detail).

Chart 1

NICE Appraisal Summary



Case Study—The Incarnation and Evolution of NREP: What can be learned for a Center for Evidence Based Medicine?

The National Registry of Effective Programs (NREP) was developed by the Federal Center for Substance Abuse Prevention (CSAP) in 1998 in response to a call to qualify the effectiveness of national demonstration grant programs. In 2002, the Federal Substance Abuse and Mental Health Services (SAMHSA), the parent agency to CSAP, adopted the NREP program for review of substance abuse treatment and mental health programs, and substance abuse prevention programs. NREP served as the research review component of a larger national dissemination system. After being reviewed, information on higher ranked programs was disseminated through multiple channels designed to enhance awareness and promote adoption of NREP-reviewed programs. Those channels included, for example, 1) posting of technical language and plain language materials on a website, 2) providing a toll-free line for researchers and consumers to call with questions concerning NREP reviewed programs, 3) direct promotional activities to create awareness of practices that have demonstrated effectiveness with specific populations, and, 4) national partnerships with key stakeholders as a primary channel for credibility and guiding dissemination strategy.

Dissemination proved to be quite effective. The number of individuals trained in specific programs increased from 1,800 in 2001 to 3,000 in 2002. Likewise organizations (train the trainer centers) increased from 2,900 to 9,900. Total number of individuals impacted by evidence-based practices increased from 1.2 million to 12.9 million. Web site hits for evidence-based practice information grew over the same time from 300,000 to over 1.0 million hits per month. Service providers and practitioners came to depend upon the National Dissemination System to keep them up to date with objective information to help them discern which evidence-based practice was most appropriate for their population (such as dosage requirements, special training of staff, cost, and outcomes).

After 2005, SAMHSA created a new direction for NREP. The dissemination of reviewed programs was discontinued and the NREP review criteria were revised. There was no clear justification for these changes. Indeed, the substance abuse field has expressed concern that the new direction of NREP does not best serve the public interest.

Why did SAMHSA drift off course with regard to NREP? 1) SAMHSA failed to provide clear policy guidelines for NREP and the National Dissemination System during the start up phases. This lack of policy guidance, embracing the work of the organization for the betterment of those disabled by substance abuse and mental illness, produced an entity that lacked sustainability during organizational and personnel changes. A change in leadership lead to a change in focus for the organization without a reasonable justification. 2) SAMHSA's failure to continue to seek stakeholder involvement (practitioners, intermediaries and consumers) in the evolution of NREP and the National Dissemination System choked interest and investment.

The NREP experience suggests that a Center on Evidence-Based Medicine needs to have a clear mandate from the beginning with guidelines that stakeholders accept. Since NREP was part of SAMHSA with no specific independence, its promising start was sidetracked into another purpose, losing its credibility and usefulness over time.

A model of a national dissemination system for evidence-based medicine has five principle components: 1) practice identification and review, 2) information development and dissemination, 3) training and technical assistance, 4) practice adoption, 5) clinical outcome review.

Practice Identification and Review

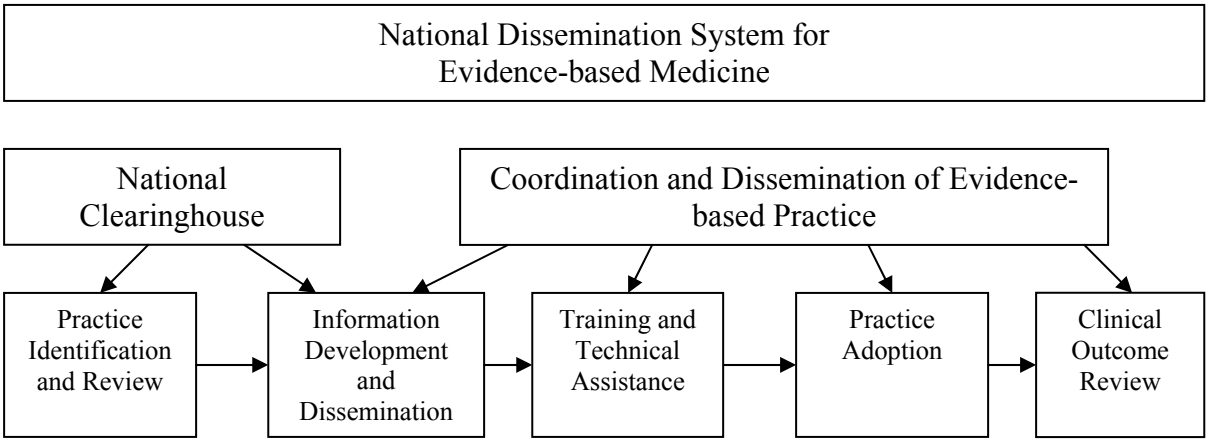
First, a national clearinghouse of evidence-based medicine would be needed to serve as a repository assisting healthcare professionals and the public become aware of scientifically defensible treatment practices. This could be part of a new organization (as is the case with NICE and NREP) or be under the control of a separate entity. AHRQ currently funds and oversees a national clearinghouse of practice guidelines. There is no reason that this could not serve that function and remain housed at AHRQ. From this clearinghouse the new center could draw studies for further rigorous review. Certainly not all research will be chosen for such review, so the clearinghouse itself could remain as a broader-based collection of research and guidelines. This provides an additional rationale for separating the more general clearinghouse function from the work of the Center. At the Center, teams of reviewers would rate interventions on a set of criteria that recognize the tenets of the research design and clinical outcomes. The team of reviewers would develop consensus scores to rank interventions

Determination of priority for research and practices reviewed is a key concern. It is reasonable to consider new or emerging interventional procedures as an appropriate starting place, or existing practices where the evidence is particularly strong and unlikely to be contested. Drug reviews are also often suggested as a potential starting point. As the credibility of the center grows, its reach could expand to a variety of medical areas.

Initially, considerable time would need to be devoted to the specific methodologies that would be used. Buy-in from the many groups actively engaged in evidence-based research such as the Evidence-based Practice Centers funded by AHRQ would be crucial. Lessons from the work done by NICE and other organizations should also be incorporated. These well-established groups need to be considered part of the process; there is no need to reinvent mechanisms that are now working well. And consensus from the research community is essential to establish the center's credibility.

Medical practitioners have competing sources of information from which they may choose to adopt a particular practice or intervention. For example, the Cochrane Collaboration conducts systematic reviews (meta-analyses) of research in healthcare; AHRQ's Evidence-based Practice Centers synthesize current evidence by reviewing published research; MED: Medicaid Evidence Based Decision Project provides state Medicaid programs high quality clinical evidence to support benefit design and coverage decisions, and DERP: Drug Effectiveness Review Project. While not an exhaustive list, this illustrates the diversity of informational sources available to practitioners. A comprehensive center would build upon the foundation established by these organizations among others—synthesizing and leveraging their work—to achieve a comprehensive system of review. A trustworthy and credible system must be based upon integrity, independence, and transparency. Thus, this system would serve the role of honest broker in qualifying evidence-based medicine. NICE uses a multi-faceted process of reviews and meetings with stakeholders before bringing forth a guidance (see Chart 2).

Chart 2



This graphic illustrates the principle components of a Center for Evidence-based Medicine. The system is comprised of two functional components 1) National Clearinghouse of Evidence-based Practice, and, 2) activities for evaluating and disseminating findings. The five principle components that constitute the activities of the system are also presented. Arrows denote responsibility and logical process flow.

An additional outcome of this work would be the standardization of the level of scientific rigor necessary to constitute efficacy of interventional procedures. Researchers and manufacturers of products and procedures would correspondingly adjust their methodological approaches to meet expectations of the center's reviewers. This will also enhance the transparency of the research process.

Dissemination

Information developed from the reviews and its subsequent dissemination must be of the highest quality, developed with guidance of key stakeholders and consumers, able to withstand vigorous scrutiny, and able to reach multiple audiences of varying levels of sophistication, in culturally appropriate and consumer friendly ways. Such comparative information may include intervention protocols, procedure fact sheets, web-based guidelines, expected clinical outcomes,

and so on. NICE, for example, produces its findings in 1, 3, and 25 page documents, often written by medical journalists. The challenge is to accurately convey the results in plain language and be viewed by stakeholders as valuable sources of information. NREP developed a considerable following in the substance abuse community which found the information helpful in thinking through changes in practices that they wished to undertake.

Dissemination should not be treated as a minor activity to be undertaken after the review process is completed. Rather, it needs to be recognized as a crucial piece of improving healthcare by reaching and convincing key audiences of the validity of the information. Behavior change that results from voluntary adoption of best practices is crucial. Even though NICE works directly with the National Health Service, it is sometimes criticized for its lack of effectiveness in reaching out to healthcare practitioners. An effective system in the U.S. with our know diversity in practice patterns would require a generous budget for dissemination.

Changes in IT, such as electronic records adoptions, could serve as a timely way to link dissemination of best practices to provide real time information to providers treating patients. This is just one example of how this organization could work constructively to improve healthcare in the U.S. in ways that go beyond simple evaluation work.

Training and Technical Assistance

Training and technical assistance are essential for clinical protocol implementation fidelity and quality improvement. Implementation fidelity enhances desired outcomes and helps to provide a standard for quality improvement. This new center could help to set up the process by developing standards for training and technical assistance. For example, it is reasonable to consider utilizing the existing network of AHRQ's Evidence-based Practice Centers to serve as

training and technical assistance (T/TA) hubs. These T/TA hubs could facilitate train-the-trainer programs, host key stakeholder groups, and provide feedback to the center on practice implementation, fidelity and adaptation issues that may influence clinical outcomes.

Training and technical assistance can take many forms including face-to-face meetings, video and tele-conferences, or seminars. The goal remains the same—utilize the most appropriate means to foster widespread adoption of evidence-based practices while supporting fidelity of practice implementation to ensure highly predictable clinical outcomes. Training and technical assistance may not be a direct responsibility of the entity but the entity needs toy provide oversight to this important activity and closely collaborate with the groups. Too often, activity such as that proposed here becomes nearly an incidental piece of the process, but acceptance and adoption of findings requires that there be a strong buy in from the health community; using only sticks in the form of coverage or reimbursement decisions would likely provide a backlash without considerable effort in this area. NREP's experience was that such efforts were useful in enhancing the esteem with which the organization was originally held.

Practice Adoption

Improved medical outcomes and enhanced service quality can only be achieved if practitioners adopt highly-rated practices. The involvement of professional associations, schools of medicine, payers and other key stakeholders as avenues of dissemination is critical to widespread voluntary practice adoption. These groups, among others, provide the necessary buy-in and network for dissemination. They also provide needed guidance for information and material development (training and guidance documents). Voluntary adoption of highly rated practices is contingent upon 3 critical issues: 1) credibility of the entity conducting reviews, 2) stakeholder involvement

and support, and 3) viable economic incentives. These issues are interrelated and serve to bolster voluntary adoption of practices.

Since economic incentives often represent a threat to opponents of creating a centralized organization of the type described here, it may be desirable to leave that implementation to policy makers rather than treating it as the role of the organization. This is the strategy used by NICE. Nonetheless, tracking of dissemination would be an important activity of the Center, and would provide information on the success of the adoption of best practices in the field.

Clinical Outcome Review

Continuous quality improvement is predicated on the process of outcome review and reconsideration of emergent scientific advances in practice areas. Although potentially controversial, one can image that over the course of time recommendations of the entity based on sustained evidence from ongoing clinical outcome reviews, may become policy. Most notably, and of particular importance, would be the discouragement of particular clinical practices that had previously enjoyed favor. Nevertheless, it is essential to the public good and the advancement of medical practice and well being of patients to maintain a review process that considers improvements or declines in clinical outcomes as essential data in the modification of treatment protocols.

Review panels would be comprised of national and international experts with special topical expertise for the intervention or practice under review. They should be carefully selected to avoid both real and perceived conflicts of interest. Many reviewers likely have conducted primary research in public and private research settings. Review panels would require extensive training in review process and protocol. Each review panel member would read, critique and

quantitatively score for each methodological criterion used to evaluate the rigor of the primary research. Panel members would then provide a detailed narrative justification for ratings. The Panel would prepare and abstract and summary addressing the research data, ratings, and narrative comments assessing strengths, weaknesses, and outcome findings. In the spirit of integrity and transparency, all materials and reports would become public documents.

THE STRUCTURE OF THE ORGANIZATION

A critical issue is the level of authority, structure, and funding such an entity would possess. It is difficult to identify an existing organization that could support a Center for Evidence-Based Healthcare, although several entities are likely better models than others. The Center should have visibility to highlight its importance, credibility in the eyes of both the research and healthcare communities, and independence from political influence. Most of the existing organizations would have difficulty assuring all three. Housing such a Center as part of HHS is likely to meet none of the criteria listed above. As part of NIH or AHRQ, for example, the Center would not stand out as prominently as desirable; it would likely compete for resources and become involved not only in political issues, but also in inter-agency struggles. The example of AHRQ at the end of the 1990s and of NREP can stand as cautions against housing a Center within a larger agency. Such agencies have other key activities to oversee and should not have the Center as simply one of many activities.

Nonetheless, the heads of NIH and AHRQ are generally noted national experts and should be participants in advising the Center and perhaps in appointing its leaders. In addition, key activities of these two organizations are worth studying for elements of the work of the Center. AHRQ not only funds the Evidence Based Policy Centers that undertake substantial

amounts of research in the United States in this area, but also maintains a Guidelines Clearinghouse described below. The NIH is actively engaged in generating consensus conferences and in dissemination of information. Careful coordination of a Center with these activities would be crucial.

A quasi-independent organization such as the Institute of Medicine is a more reasonable model for the Center. Again, the Institute of Medicine has a broader agenda and the Center should not simply be tacked onto this organization. But the charter for the Institute of Medicine and the role it often plays in tackling sensitive subjects to aid the public debate on key healthcare issues such as quality and patient safety suggest that this is the type of independence that can tackle controversial work and come up with clear consensus statements. The IOM is also able to accept funding from other sources such as foundations to extend the reach of its work. However, the IOM often must scramble for funding and depends on study panels of volunteers—both of which could hamper the smooth operation of an organization that will need multiple years of stable funding in order to establish its reputation and undertake multiple studies every year. While the size of the staff could be relatively small, it will be important to have sufficient funding over a period of years rather than competing for annual appropriations.

Funding should likely come from a source that allows long term commitments rather than being subjected to annual appropriations processes. Given our fragmented healthcare system, it might also be desirable to have funding supplemented by other entities such as national foundations and perhaps a consortium of other payers (although careful controls on interference would be necessary). Funding should also be sufficient to allow some basic research to be supported to fill in gaps, but most research would not be funded by the center. A period of 3 to

5 years may be necessary to implement and nurture to maturity the credibility and key stakeholder involvement to make this venture viable.

As envisioned here, the funding for the Center would not have to be that large. A small permanent staff, visiting scholars, and advisory board members would not constitute a large bureaucracy, but rather a small organization drawing on other sources for the expensive activities surrounding research and large-scale dissemination.

Appointment of leadership for the organization ought to occur in a very open manner with the goal being well-respected authorities in evidence-based research. Professional staff members need to be top-notch. The positions should be sufficiently desirable to be able to attract outstanding scientists, economists, and other professionals. One way to foster excellence might be to have some of the staff comprised of visiting professionals, taking a two year sabbatical at the Center, for example. The goal from the beginning should be to make this organization a coveted workplace. In addition to those doing the reviews, it will also be important to have professionals in dissemination and plain language writing as part of the organization.

At least two advisory panels would need to be part of the Center. First, a panel made up of distinguished scientists, researchers, and practitioners need to assess each "guidance" developed for release. Given the broad range of potential topics, the panel would likely be quite extensive with subgroups meeting on specific topic areas. A second advisory panel of stakeholders, patients, and communication experts should also be involved in the development of materials for dissemination. It is important to assure buy in from various groups; debate on the science is critical, but so is an assessment of the practical considerations that will arise from the findings in controversial areas.

As has already been mentioned it will be important not to replace existing efforts that are already making strong contributions. Where possible, the Center should be viewed as part of a broader effort to provide and facilitate evidence-based practice. Much of the research that is already underway in the U.S. and elsewhere would feed into the process; the Center would likely only recommend funding to fill in important gaps. And at the decision-making end, the Center should be similar to NICE which turns its findings over to others for implementation decisions. It is essential that the Center not be simply viewed as a part of the insurance or coverage process, but rather an integral part of health education, research and practice in the United States.

KEY ISSUES IN ESTABLISHING A CENTER FOR EVIDENCE BASED RESEARCH

The structure of the organization represents only a part of the decisions that need to be made in establishing an effective center for evidence-based research. The general goals of the center, relationships with other potentially competing groups, the independence of the organization and its approach to undertaking its work are all critical elements upon which there needs to be agreement from the very beginning.

The Goals of the Organization

The mandate for this type of entity should be much broader than simply assessing the quality of evidence-based research. The outcomes to be desired from the organization should be broader—essentially to foster improved healthcare and resulting outcomes for the population. This goes well beyond any charge of cost containment or information sharing, for example. If it is viewed as having a broader mandate, the expectations of the organization will be for an inclusive and forward-looking framework. It is also likely to be easier to gain buy-in from stakeholder groups if it has a range of activities, some of which are viewed as building on ongoing work.

At the same time, it is important not to raise expectations too quickly. An organization of this type must have the time to build its constituencies and audiences and establish trust across a wide array of groups with sometimes competing goals. The medical establishment must be on board with the goals and the activities that the organization will undertake. Patients' views and concerns need to be an essential part of the endeavor as well. The organization will stand a greater chance of success if it is viewed as helping to improve healthcare rather than simply enforcing limits or bounds. Voluntary adoption of findings would be the best outcome of the work of such a group. That would mean that any ultimate decision-making or enforcement activities would only have to deal with outliers rather than with changing the behavior of the majority of health care providers.

Essential Elements

People familiar with NICE and NREP stress that political challenges will arise almost immediately. There are a few key ways to maximize chances for success:

- The process of conducting reviews must be viewed as reflecting consensus based on the best possible evidence. The more that experts in the field can be used to validate the methodologies and ratings, the better.

- Transparency in the process is also critical. The standards for reviews must be clearly spelled out and explained in the findings. Challenges will be launched and must be addressed head on. If errors are found, they must be acknowledged and corrected immediately.

- The organization must have and maintain its independence. Stakeholders should be allowed to submit their own findings and evidence, but all work needs to be subject to the same standards for review. Political interference needs to be minimized, likely by making the entity quasi-governmental with long term funding (for example, from the Medicare trust funds) not subject to annual appropriations processes. The organization must have on staff individuals who are knowledgeable about the political issues and environment in which this organization will have to operate.

As has been mentioned throughout this paper, a center of evidence-based healthcare will be isolated and ineffective unless it is embraced by key stakeholder groups. This includes providers, payers and the general public. It will be facing an uphill battle unless it is viewed as an independent and credible authority providing valuable information. Some further ways to obtain stakeholder buy-in include:

- Make sure that thought leaders are included in the process, particularly in developing agreement on the methodology to be used and maintaining high standards for the qualifications of those doing the analysis.

- Allow for review and adjustment of findings as new results become available over time.

- Focus on the strongest findings initially and allow a finding of "not enough evidence" to occur when the studies are equivocal.

- Tie these efforts to rewards for excellent research or to highlight new breakthroughs. This should be a "good news" organization as well as one that discourages use of certain drugs or procedures.

- Spend time with stakeholders in the initial phases of the project and develop partnerships when that makes sense. Rather than being exclusive, this organization should be seen as a source of fair brokering of disputes.

- Do not reinvent the wheel. This organization should build on others' work, not seek to displace it.

- Take care to avoid real and perceived conflicts of interest.

CONCLUSION

Like many promising areas for improving health care, the likely impact of evidence based research is often oversold. Indeed, it would be difficult for any new organization simultaneously to raise the quality of evidence-based research, alter the way in which care is delivered, reduce variability around the country, and reduce costs at the same time. Yet such are the gains sometimes claimed from the application of evidence based work. The first three should be the initial mandate for a new national organization seeking to establish a credible, constructive role in improving healthcare.

APPENDIX

EVIDENCE-BASED MEASURES

Analyses of health care goods and services run across a broad spectrum of evaluation research, and what constitutes "evidence" is not always well defined (Steinberg and Luce 2005). The present emphasis on measuring healthcare outcomes is relatively new. Abraham Flexner's (1915) treatise on training physicians signaled a shift from viewing medicine as something of a magical art to seeing doctors as science-practitioners guided by performance and ethics standards. Thus, socializing doctors into normative roles became the primary goal of physician training. Despite this early 20th Century shift, however, patients continued to defer to doctors as the ultimate arbiters of medical decisions.

Basic effectiveness studies on a single drug or treatment have been in place and used for decision making for many years. For example, the Federal Drug Administration requires that pharmaceutical manufacturers wishing to market their drugs in the United States must provide controlled trials demonstrating efficacy and safety. Moreover, Medicare has strengthened its reviews of new technology and treatment through its Medicare Coverage Advisory Committee.

Stakeholders wishing to obtain coverage know that they must demonstrate the efficacy of new treatments and drugs. It is not surprising that this is not a controversial area because approval means that the array of options available to providers of care expands over time with only the requirement that a positive benefit/risk ratio must be documented. Critics of the U.S. healthcare system note, however, that this does little to reduce unnecessary care or address issues of over-use of services. In some cases, it may even encourage unnecessary use of services.

Often, the next step for moving to more comprehensive evaluation research is undertaking comparative effectiveness analyses in which two drugs or two treatments, for example, are contrasted with each other to try to determine whether they are equivalent or whether one is superior to the other. These head-to-head comparisons have been much less common in the U.S. but are routinely used elsewhere. For example, in Australia, a manufacturer wishing to have its drug listed on the national reimbursement formulary must demonstrate that it is at least as efficacious and safe as the most commonly prescribed therapy for some condition. For example, a new cholesterol lowering therapy would need to establish that it is not more toxic, and is at least as efficacious in lowering cholesterol as one of the most commonly prescribed "statins." Where formularies require such comparative analyses, these may be used to limit the number of "me too" drugs available, or where a degree of choice is valued, to create competition between treatments while ensuring that drugs for which there is no evidence of an improved benefit do not attract a higher subsidy. In the U.S., this higher standard is not required by the FDA in the regulatory approval process, nor by other groups that evaluate treatments, procedures or devices. Introducing new criteria into decision making is always controversial, but particularly in this case, because this type of approach has the potential for *reducing*, not expanding, options available to providers and consumers of care. And it is exactly this aspect of comparative analysis that makes it appealing to some and threatening to others.

To traditional medical researchers the research design known as the randomized clinical trial remains the *sine qua non* of research methodologies. As such, we would expect randomized clinical trials to account for many of the findings that an evidence-based center might propose to advance. For example, the main purpose of the Cochrane Collaboration and the Evidence-Based Practice Centers (funded by AHRQ) is to evaluate peer-reviewed research by scoring studies

according to a strict set of criteria. Study results that meet established standards can then be incorporated into subsequent meta-analyses. Both the Cochrane Collaboration and the Evidence-Based Centers routinely award a higher score to randomized clinical trials than to any other design, including observational and large-scale epidemiological studies. Given that the randomized trial is the cornerstone of the evidence-based movement, its strengths and limitations merit attention.

The major strength of the randomized clinical trial is that, assuming a sufficiently large sample, it eliminates systematic bias. That is, because the researcher administers a treatment before he or she assesses its effect, the results imply a causal chronology. In fact, while such findings are necessary, they are not sufficient to establish causality. In addition, randomly assigning participants to treatment arms ensures that the potential effects of any extraneous variables will be distributed randomly across conditions and thus controlled.

Although the randomized clinical trial has served as a mainstay of healthcare research, it has drawn considerable criticism, especially in its most ubiquitous form. Specifically, randomized clinical trials protocols generally limit the selection of subjects to those within a certain age range (i.e., they will tend to exclude the very young and the very old) and with only the medical condition of interest (e.g., diabetes). However, many individuals who comprise the population in whom the treatment will ultimately be used will have multiple co-morbidities and may be older or younger than the subjects in the RCTs. In such cases, generalizing the study's findings from the study subjects to the wider population may not be straightforward; the study's results demonstrate internal validity, but not external validity (i.e., they may support a conclusion about "efficacy" but not of "effectiveness"). Recognizing the prevalence of this limitation, researchers charged with evaluating healthcare effects increasingly emphasize the need for

randomized clinical trials that test participants who exhibit a cluster of co-morbid conditions and/or who take more than one medication (Cochrane Handbook, 2006; Miller, Robinson, & Lawrence, 2006).

Drawing unjustified inferences regarding external validity is particularly likely to occur in medication trials sponsored by pharmaceutical companies. Because substantial profits are at stake, companies may select a study sample and a comparator therapy and utilize outcome measures designed to maximize differences between treatments. For example a study might compare an arthritis medicine to ibuprofen when the arthritis medication has been designed not to cause stomach problems and ibuprofen is a proven stomach irritant. Then, if the outcome measures were weighted towards observing stomach-related side effects—especially at the expense of identifying other complications and determining the superiority of the new drug for arthritis treatment per se—the results would be misleadingly optimistic.

Another limitation of randomized clinical trials is attrition, especially if uneven attrition occurs across treatment arms. Although researchers tend to think of these trials as "experiments," they are, in fact, "quasi-experiments." Except when they are present at the treatment site, participants lead their regular lives. Thus, clinical trials are vulnerable to all the biases that can plague other quasi-experimental designs—participants' previous experience, events that occur to one group but not to another during the experimental interval, participant attrition, knowledge of the other group's experience, etc. (Cook & Campbell, 1979). In short, even in the area of RCTs, rigorous review is likely to be necessary in rating the strength of results.

Taking analysis a step further to examine the comparative cost effectiveness of healthcare means that the issue of the value of the additional benefit obtained from a new drug or procedure should exceed its costs. Both value and costs can be broadly defined, making these analyses even more controversial; the implications for restricting use (where the cost exceeds the value), along with the lack of full acceptance of the methodological approaches and potential variation and uncertainty that can arise in such studies, both contribute to the controversy. Nevertheless, many researchers consider cost-effectiveness analysis essential to improving healthcare treatments and delivery.

For those advocating cost effectiveness studies, four tacit assumptions underlie contemporary healthcare-evaluation research: first, that valid measures of healthcare effects (both positive and negative) and healthcare costs can be developed; second, that a ratio of treatment costs to treatment effects can provide meaningful information about the worth of a treatment—to an individual, to a population, or to a society; third, that a ratio of the differences between the costs and effects of two treatments can reveal their comparative worth; and fourth, that such comparisons can inform public healthcare policy. For the last three of these assumptions, currently used "effects" measures derive from economics models, decision theory, and operations research (Gold et al., 1996). Effects measures related to a single treatment may also reflect disciplines like sociology and psychology.

With the general acceptance that the most useful models for healthcare evaluation would come from economics and with the growth of the patient-centered-care movement, the medical zeitgeist began to shift again in the 1980's. Since then, the emphasis on quantitative assessments of health costs and effects has increased steadily, especially over the last decade. As a result, both regulatory requirements and standard research practice now mandate quantitative, codified

outcome measures for virtually every type of service. By offering a methodology for generating quantitative evaluation data, cost effectiveness analysis addresses this mandate.

Cost effectiveness analysis involves stipulating a ratio that compares treatment costs to treatment effects. Two types of comparative analyses—both of which have been used in evaluation research and to guide public health policy—are based on cost-effects ratios: benefit-cost analysis (BCA) and cost-effectiveness analysis (CEA). The difference between BCAs and CEAs is that benefit-cost analyses define both costs and benefits in monetary terms; in contrast, cost-effectiveness analyses use monetary units only to define costs (Gold, Siegel, Russell, Weinstein, 1996; Miller et al, 2006).

Government regulations have traditionally required agencies to conduct benefit-cost analyses that assess the efficacy of the public health and safety interventions they implement. Given that the BCA's numerator and denominator are expressed in the same unit of measurement, these analyses have appeared to facilitate straightforward comparisons between two treatments or between a treatment and no treatment.

However, the measurement units of the benefits that policy-makers and researchers alike find most relevant to treatment comparisons—uni-dimensional measures like the number of deaths averted or multidimensional measures that integrate longevity with quality of life—are virtually impossible to define monetarily (Gold et al., 1996; Miller et al., 2006). In recognition of this limitation, the Office of Management and Budget (OMB) issued the 2003 requirement that all future BCAs be supplemented with CEAs (Miller et al., 2006).

Nevertheless, with respect to comparing an intervention's long-term costs and benefits, CEAs, like BCAs, can seriously underestimate a program's effectiveness. This bias results from

combining "discounted" costs with an unequal distribution of benefits over time (Gold et al., 1996). Taken from the business literature, the central tenet of discounting is that the present value of a given unit of currency is higher than its future value. Thus, applying a discounting factor of five percent per year means that next year's dollar will be worth only $.95. As a result, discounting a future benefit means that, even if the outlay of resources remains constant, the dollar cost of the intervention will rise relative to its benefits. Moreover, benefits tend to be unequally distributed because they generally take time to develop. For example, the benefits of lifestyle interventions designed to prevent heart disease or diabetes will not be observable until years later. However, when benefits do not emerge until the latter portion of an evaluation period, the cost-benefit ratio may erroneously reflect that the program is losing money, without delivering results.

As stated above, one reason that the research and health-policy communities sought a replacement for (or at least an adjunct to) benefit-cost analysis is that many of the most relevant outcomes cannot be defined monetarily. In addition to this measurement-oriented motive, however, evaluators found placing a monetary value on human life to be aesthetically and ethically objectionable (Gold et al., 1996). Together, these reasons provided the impetus to look for a new way to assess benefits—the denominator of the CEA ratio. The majority of controversy generated by cost-effectiveness analysis has involved disagreements over the appropriate method of defining this denominator.

The primary family of metrics to emerge from the search for non-monetary benefits measures have been the HALY (health-adjusted life years) measures. Among the available HALY measures, international agencies, such as the World Bank and the World Health Organization, have tended to use the DALY (disability-adjusted life years) metric (MH; Prüss-

Üstün, Mathers, Corvalan, & Woodward, 2003). However, in CEA ratios defined for comparative evaluations in the U.S., the most commonly used HALY is the QALY (quality-adjusted life years).

Basically, a QALY measure is calculated by assigning a weight, ranging from 0 to 1, to each period of time in the study. These numbers reflect the *quality of life* experienced during that period; a weight of 1 represents perfect health, whereas a weight of 0 essentially represents death. The resulting number of quality-adjusted life years is assumed to represent the number of healthy years of life that will result from the assessed health outcome (Gold et al., 1996).

Ostensibly, QALY measures not only avoid the dilemma of assigning monetary values to lives, but they offer the added bonus of integrating all relevant outcomes into a single composite. Moreover, because different combinations of outcomes can be translated into directly comparable QALYs, *interpreting* comparative evaluations is as straightforward as it would be if the CEA-ratio denominators were expressed in dollars. Finally, individual-level QALYs can be summed to represent integrated health effects at the individual, the group, or the population level (Gold et al., 1996; Miller et al., 2006).

Although cost-effectiveness analysis generally proceeds as though QALYs provide all the advantages described above, these metrics suffer from certain limitations. First, it should be noted that—masked though the process may be—assigning a QALY score still represents placing a value on human life. Hence, the use of QALYs requires researchers to exercise the same ethical concern that discomfits them when they assign dollar values to lives. Furthermore, in a standard QALY calculation, the quality-adjusted life years that represent the benefits in a particular analysis are considered to signify the same "quality," regardless of age differences,

differences in personal ability to control risk, or any other heterogeneity in the population of interest: "a QALY is valued the same …, regardless of who is affected" (Miller et al., 2006, p.8).

Applying a discounting factor to correct for such differences, though necessary to avoid bias, is comparable to discounting costs; thus, in the context of CEA, the value of life, like the value of money, decreases over time (Gold et al., 1996).

The methods used to generate QALY scores demonstrate that their primary limitation is their *own lack of validity*. The initially-established process for developing QALY measures is to use an individually administered successive- approximation procedure, to determine the exact trade-off point at which each individual is no longer willing to compromise longevity for quality-of- life. The validity of this procedure relies on a two-part assumption. The first part assumes that individuals can *imagine* how—*if* they contracted a life-threatening health risk, like cancer—they would make a decision regarding quality versus length of life. The second part assumes that this decision will remain stable, regardless of the time interval over which cancer does or does not occur.

Not only does neither part of this assumption appear to be supported by empirical evidence, but a substantial body of research challenges the fundamental assumption underlying QALY development—namely, that individuals can and do make rational decisions under conditions of uncertainty. In opposition to this belief, Khaneman and Tversky (1979) developed and validated "prospect theory," work for which Kahneman won the 2002 Nobel Prize in economics.

Prospect theory was specifically conceptualized as a psychologically realistic alternative to expected utility theory (which is the basis for QALYs). By studying how individuals choose

between risky alternatives, Kahneman and Tversky identified the cognitive processes through which humans evaluate potential losses and gains (Bernstein, 1996). As described by Bernstein (1996), Kahneman and Tversky demonstrated that individuals' attitudes toward risks involving *gains* can differ appreciably from their attitudes toward risks involving *losses*. Furthermore, depending upon the context in which a risk is offered, the same individual may either avoid it or seek it.

In addition, Bernstein (1996) notes that the problem of interpreting human behavior in the face of risk sometimes entails individuals' making decisions on the basis of incorrect probability assessments. Thus, individuals may mistakenly assign a probability of *zero* to a low-likelihood event that has not yet occurred. But because small probabilities can combine over frequent risk-taking behavior, individuals may, metaphorically speaking, find themselves engaged in a game of Russian roulette (Bernstein, 1996). In sum, the strong empirical validity that prospect theory has accumulated would seem to undermine the central premise that supports the QALY methodology.

Later-developed methods of building QALY measures, described by Gold et al. (1994) and Miller et al. (2006), seem only to aggravate the concerns surrounding the initial method. The later methodology has resulted in numerous indices, each of which provides QALY scores that reflect levels of functioning and well-being related to a particular health risk. Essentially, the scores are ready-made trade-off estimates that are assumed to be valid for the health condition being indexed, regardless of the population or contextual characteristics that might be unique to a particular intervention. The indexed scores replace the QALY data that would otherwise be gathered as part of each individual study.

In summary, although the IOM Report concludes that QALYs represent the best of the current CEA measures, Miller et al. (2006) nonetheless emphasize the need "to improve the quality, applicability, and breadth of HRQL measures for use in regulatory CEA" (p. 13-14). Given the concerns outlined above, comparative evaluations might profit from researchers' seeking non-HRQL-related ways to measure CEA outcomes.

The final type of outcome measure discussed here consists of survey responses from patients. Known as "patient-reported outcomes" (PROs), these measures are commonly weighted using psychometric methods and can be used to assess a wide array of health status domains or "PRO concepts." At the uni-dimensional level, PRO concepts can reflect symptoms experienced in relation to a specific disorder or disease (e.g., anxiety, nausea, headache); at the multidimensional level, they can represent holistic functioning, either with respect to a particular life domain (i.e., ability to carry out activities of daily living) or with regard to functioning in the physical, psychological, and social spheres that define overall quality of life.

One of the main challenges associated with developing valid PRO measures concerns determining the appropriate level of item specificity. On the one hand, the measured outcomes should be sufficiently specific to the intervention being assessed to allow the evaluator to draw inferences concerning the intervention's efficacy—that is, its success within the context of the study. On the other hand, the measured outcomes should be sufficiently broad to allow the evaluator to draw inferences concerning the intervention's effectiveness—that is, its generalizability to the population of interest. The difficulty of striking a suitable balance depends, at least in part, upon the purpose of the study in question.

For example, for PROs to provide useful information in a *comparative* evaluation, the outcome measures must, in fact, be comparable. Founded on investigations that began in the 1970's, an ongoing program of research has addressed this issue, both substantively and methodologically. From a substantive perspective, the development of PRO measures evolved from a focus on chronic diseases to an emphasis on assessing what researchers designated as the two primary health domains: functioning and well-being. This latter emphasis led to the development of the *SF-36* (Ware & Sherbourne, 1992), which is currently the most widely used PRO measure in the world. The trajectory of this substantive progression has resulted in PRO surveys (the *SF-36* and others) that, generally speaking, are no longer maximally sensitive to any one disease. Simply stated, comparability has trumped specificity.

The use of Item Response Theory to develop PRO scoring algorithms further adds to their comparability. Item Response Theory (IRT), which was first used as a replacement for classical psychometrics in analyzing cognitive-ability tests, considers the construct being measured to be an underlying "ability" continuum. For example, a functioning continuum ranges from "ability" to "disability," whereas a well-being continuum extends from "not suffering" to "suffering." Through analyzing all the respondents' scores, IRT programs place each survey item somewhere on the defined continuum. Each item's placement models the probability that an individual with the level of functioning (or well-being) reflected by that point on the continuum will answer that item in that way. Because IRT parameters are assumed to be invariant across populations, once the program maps the items onto the construct, the scores are comparable.

Potentially, however, the most significant breakthrough in PRO-measure research comes from NIH's dedication of 25 million dollars to developing and implementing a PRO

Measurement Information System (PROMIS), comprising 10–12 item banks. Because the items will be administered through Computer Adaptive Testing, individuals' responses to each question will determine which other questions are presented. As a result, each respondent will respond to a survey that has, to a large extent, been tailored to his or her unique situation. Moreover, advanced IRT methods will allow scores on different items to be equilibrated, both across diseases and across populations. Thus, the promise of the PROMIS is that the respondent's specificity will become the researcher's comparability. Of course, the extent to which this strategy will meet expectations cannot yet be determined.

As this discussion of measures indicates, there is no easy way to establish a gold standard on the validity of evidence-based research. Nonetheless, improvements in measurement, transparency regarding methodology, and building consensus on techniques can help to raise the credibility of research that will likely be increasingly important as the cost of healthcare continues to rise. This will be one of the most important roles for a center on healthcare improvement.

References

Bernstein, P. (1996). *Against the gods: The remarkable story of risk*. New York: John Wiley & Sons.

Cook, T. D, & Campbell, D. T. (1979). *Quasi-experimentation*. New York: Rand McNally & Co.

Handbook of Systematic Reviews of Interventions 4.2.6. (2006). Oxford, UK: Cochrane Collaboration.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost effectiveness in health and medicine*. New York: Oxford University Press.

Gray, Bradford, Michael Gusmano and Sara Collins. ((2003) AHCPR and The Changing Politics of Health Services Research. *Health Affairs: Web Exclusive*, June 25, W3-283 – W3-298.

Harvey, C. M. (1994). The reasonableness of non-constant discounting. *Journal of Public Economics*, *53*, 31-52.

Kahneman, D., & Tversky, A. (1979a). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 313-327.

Martin, S. L., & Raju. N.S. (1992). Determining cutoff scores that optimize utility: A recognition of recruiting costs. *Journal of Applied Psychology*, *77*, 15-23.

Miller, W., Robinson, L. A., & Lawrence, R. S. (Eds.) *Valuing health for regulatory cost-effectiveness analysis*. (2006). Institute of Medicine of the National Academies (IOM). Washington, DC: National Academies Press.

Pearson, Steven and Michael Rawlins. 2005. Quality, Innovation and Value for Money: NICE and the British National Health Service. *American Medical Association*, 294, 20, 2618 – 2622.

Prüss-Üstün A., Mathers, C., Corvalan, C., & Woodward, A. (2003). *Introduction and methods: assessing the environmental burden of disease at national and local levels*. WHO Environmental Burden of Disease Series, No. 1. Geneva: World Health Organization.