The quality of medical evidence: implications for quality of care

D M Eddy and J Billings

Health Affairs 7, no.1 (1988):19-32
doi: 10.1377/hlthaff.7.1.19
THE QUALITY OF MEDICAL EVIDENCE: IMPLICATIONS FOR QUALITY OF CARE

by David M. Eddy and John Billings

Prologue: Most of the proposals advanced in the 1980s for rationalizing the use of medical care derived chiefly from an economic model—more consumer cost sharing, the development of alternative health delivery systems, and new financial incentives. One of the working assumptions of the analysts and policymakers who designed these proposals was that the scientific basis of medicine was solid. If purchasers of care could only determine how to make providers more efficient, they reasoned, the growth of medical costs could be moderated. In this paper, David Eddy and John Billings call into question that assumption. They make a case for why it is imperative that the medical profession get on with the job of examining more closely the scientific basis of medical practice. Eddy holds a medical degree from the University of Virginia and a doctor of philosophy degree in applied mathematics from Stanford University. Eddy, who currently directs the Center for Health Policy at Duke University, is the J. Alexander McMahon professor of Health Policy Research and Management. Eddy operates at the intersection of math and medicine, applying probability theory to the uncertainty of approaches to care. Billings holds a law degree from the University of California, Berkeley, and has been actively involved in a variety of health care activities for a decade. Billings, who is currently an independent consultant, wrote this article with Eddy while serving as a visiting professor at Duke. He served as one of the first staff directors of a community-based (Salt Lake City) health care coalition (1979–1980) and was executive director of The John A. Hartford Foundation for four years (1981–1985).
To achieve high-quality medical care, we must succeed at three main tasks. First, we must determine just what practices constitute high-quality care. This involves analyzing evidence of the effectiveness, risks, and costs of various medical practices, and designing standards that define appropriate practices.\(^1\) The second task involves monitoring existing practices to compare them against the accepted standards. The third involves changing the behavior of practitioners to ensure that the care actually delivered meets the standards. Failure at any of these tasks will threaten the quality of care people actually receive.

This article examines our current ability to perform the first task. It examines the quality of evidence and the quality of analysis that supports some current standards of practice. It will argue that, for at least some important practices, the existing evidence is of such poor quality that it is virtually impossible to determine even what effect the practice has on patients, much less whether that effect is preferable to the outcomes that would have occurred with other options. Furthermore, whatever the quality of the existing evidence, our current ability to analyze that information is primitive. As a consequence of these two findings, we simply do not know the appropriate standard of care for some medical practices. The care that is currently being delivered might or might not be appropriate. The standards we use to evaluate actual practice may or may not be the correct ones. The article concludes that, to design truly effective quality assurance programs, we must do considerably more work to obtain better evidence about the health and economic effects of different options and to use that evidence in designing appropriate standards of practice. Accomplishing this will require not only specific actions but also a consolidation of leadership, the development of a common vision and set of principles, and coordination and compromises in executing those principles.

**The Problem**

Determining whether any particular practice is appropriate requires two main steps (Exhibit 1). First, the available empirical evidence must be evaluated to estimate the effect of the practice on health outcomes. (We use the term “health outcome” to describe an outcome of a disease or injury that a person can experience and that affects a person’s length or quality of life. Examples are death, pain, anxiety, disability, and disfigurement.) Because the available evidence is never complete, this invariably requires some use of clinical judgment. The second step is to compare the benefit and harm (for example, risks and side effects) of the practice to determine whether, in the eyes of the people who will have to live with the outcomes, the benefit outweighs the harm.

It is also important to estimate the resources required (such as costs) to
determine if the health outcomes are worth the resources, and, if resources are limited, to set priorities. This article will focus only on estimating and comparing a practice’s health outcomes.

**The Quality Of Medical Evidence**

Ideally, for each practice there will be direct evidence from one or more studies that relates the use of the practice (compared with a specified alternative) to the health outcomes of interest to patients. Types of direct evidence include randomized controlled trials and case-control studies that observe the health outcomes important to patients. The discouraging fact is that the direct evidence that exists can be so poorly designed and reported that it is virtually impossible to estimate from it the effects of a medical practice—how it will change the probabilities or magnitudes of the outcomes that are important to people. The problem is illustrated by an examination of the evidence for percutaneous transluminal angioplasty (PTA) and a comparison of PTA with an alternative treatment, bypass surgery.

The literature on PTA was analyzed recently by Raphael Adar, chief of general and vascular surgery at the largest medical center in Israel, during a visit to Duke University’s Center for Health Policy Research and Education. He found thirty-nine papers that addressed the use of PTA in the leg. In our experience, this is not unusual; for most of the procedures we research, a large number of papers purport to study its effects. Unfortunately, most of the papers were not useful, which also is not unusual. In this case, twenty-seven had to be excluded for one reason.
or another, such as mixing the results of treating patients with vascular disease of the leg (the femoral artery or popliteal artery) with the results of patients treated for vascular disease of the pelvis (the iliac artery). These are two distinctly different anatomical locations, and we can expect the effectiveness of treatments to be different in each location. When the results are combined, it is not possible to determine the effectiveness of the procedure for either group of patients.

There is a serious problem with every one of the remaining papers. The worst is that not one study was controlled, much less randomized. That is, in none of the studies were patients randomly assigned to receive either PTA or bypass surgery. Without this, it is virtually impossible to tell what the effectiveness of bypass surgery would have been, or how much better or worse PTA is compared to surgery. A second problem is that more than half the papers mixed the results of PTA for patients treated for different indications. There are two main indications for treatment, and the preferred treatment can be different for patients with different indications. Some patients have pain on walking (intermittent claudication) but can still use their legs. Other patients have such extensive disease that the survival of the leg is threatened, and the main purpose of treatment is to save it ("salvage"). Because the severity of the disease is so different in these two groups of patients, we can expect the outcomes of PTA to be different. By mixing the two types of patients, it is not possible to determine the effectiveness of PTA in either patients with claudication or patients treated for salvage. A third problem is that none of the papers reported the outcomes most important to patients: relief of pain or ability to walk. Rather, they reported "intermediate" outcomes that are easier to measure but that are only indirectly related to the outcomes of real interest to patients—successful opening of the artery ("immediate success"), or how long the artery stays open ("long-term patency"). Fourth, even for these outcomes there was no common format for reporting results. A fifth problem is that none of the papers reported all of the important outcomes. Sixth, even if these problems could somehow be corrected, additional problems would have to be resolved; the studies differed with respect to the types of patients included, the techniques used, the skill of the surgeons and radiologists, and so forth. Finally, an analysis of the literature on bypass surgery does not help. It also consists of uncontrolled studies, with its own list of problems.

The only possible conclusion is that, given the available evidence, there is no way to determine with any degree of accuracy the relative merits of the two approaches. How then do practitioners decide which procedure to use when patients enter their offices? It is impossible to say for certain, but it is clearly not on the basis of empirically based estimates of the effectiveness of the two approaches, because this is unknown. We
can only surmise that the procedure chosen might depend on which specialist dominates the decision. A nonsurgeon is likely to choose PTA; a surgeon, bypass surgery.

This example illustrates the type of problems that can exist. The following are clues that the problems might be widespread. (1) The Congressional Office of Technology Assessment has estimated that only 10–20 percent of practices are supported by randomized controlled trials. (2) A review of the statistical methods in published articles concluded that “approximately half the articles that used statistical methods used them incorrectly.” (3) A group of published papers about the effectiveness of magnetic resonance imaging (MRI) were scored for quality of design and reporting; on a scale of 1 to 100, the average score was 13. In short, a major problem relating to the first task of a quality assurance program is that for some practices we simply do not have the direct evidence needed to determine the outcomes of the practice.

## The Quality Of Analysis

When there is no direct evidence about the effects of a procedure, people who make decisions or design standards must build a case from indirect evidence. They must piece together information about a variety of factors relating to the biology of the disease and the mechanism of action of the intervention. This raises three additional problems. First, analyzing a health problem can require digesting a large number of factors and options—sorting it all out can easily exceed the capability of the unaided human mind. Second, even for a relatively small piece of a problem, there can be many factors. Synthesizing evidence about even these factors can be extremely complicated. Third, the evidence about even one important factor might not be very good.

All three problems are illustrated by screening for colorectal cancer, a disease that causes about 145,000 new cases and 60,000 deaths a year. Many national and international organizations make recommendations about how people should be screened for this disease. Here are some of the factors they must keep in mind: (1) Invasive colorectal cancers can develop from at least two different sources (adenomatous polyps, or more simply adenomas; and de novo). (2) Adenomas and de novo cancers have different natural histories (for example, growth rate, propensity to bleed) and develop in different regions of the intestine in different proportions. (3) There are at least four major types of screening procedures—digital rectal examination, fecal occult blood test (FOBTs), endoscopes (lighted tubes inserted through the rectum), and barium enemas. (4) For three of these types there are several options, with about a dozen brands of FOBTs; rigid or flexible (fiberoptic) endoscopes that come in at least four different lengths; and two basic types of barium
enemas (single or double contrast). (5) Each procedure can reach different areas of the large intestine, and has a different sensitivity (probability of finding an existing cancer) and false-positive rate (probability of being positive in a person who does not have cancer). (6) The tests can be used in any combination, at any frequency, starting and stopping at any age. (7) Different strategies can be used for people with different risk factors. And so forth.

To appreciate the complexity of analyzing even one piece of this problem, consider the FOBT. This screening test is based on the fact that occult (or hidden) bleeding found in a stool sample can be an early sign of an invasive cancer or adenoma. The test is usually recommended for people fifty and older on an annual basis; if such recommendations were actually followed, this would represent about 62.5 million applications a year. Clearly, for any procedure intended for 62.5 million people, we want to know how the test will affect important outcomes. In this case, the most important outcomes are how screening with FOBTs will decrease the chance of ever getting colorectal cancer (by finding and removing an adenoma before it becomes an invasive cancer) or decrease the chance of dying of colorectal cancer (by finding an invasive cancer in an earlier stage when treatment might be more effective). No completed controlled trials have yet published direct evidence on these questions, and the recommendations in place today had to be based on indirect evidence.

Just what does such evidence look like? Consider one of the questions—how FOBT screening might change the chance an individual will die of colorectal cancer. There are three main factors to be considered: the chance an individual actually has an adenoma or cancer that might possibly be detected; the chance that, if a person has such a lesion, the FOBT will detect it (the sensitivity of the FOBT); and how detecting a lesion through screening (before the onset of sign or symptoms) will affect survival. In turn, each of these factors is affected by additional factors. For example, the probability an asymptomatic individual is harboring an adenoma or cancer is determined by (1) the annual incidence rate, which is a function of the person’s age, sex, and risk factors; (2) the growth rates of adenomas and invasive cancers; (3) the time of the last screening examination, if any; and (4) the probability a previous screening examination had a false-negative result (that is, a cancer or adenoma was present but missed).

All of these factors must be considered just to understand the effectiveness of screening with the FOBT. A large number of additional factors must be considered to address other important questions relating to the use of other screening procedures, the choice of an optimal age to start screening, the age to stop screening, the frequency of screening, and the design of different protocols for different risk groups.
If this is confusing, that is the point. Synthesizing indirect evidence is extremely complicated. It is neither a surprise nor an insult to say that even when good information is available on all the important factors, putting them together can easily exceed the capacity of the unaided human mind. Mathematical models can be used to structure the existing knowledge and perform the calculations, but the overwhelming majority of problems are analyzed without such models.

Unfortunately, the problem of analyzing indirect evidence is compounded by the fact that we often do not have good information on all the factors. The complexity of the problem and the poor information about various factors leads to what you would expect. There is an extremely broad range of uncertainty about the actual impact of colorectal cancer screening on the most important outcomes, whether someone will get colorectal cancer, and if they get it, whether they will die from the disease. A group of international experts that met periodically to review the evidence and design screening policies was asked to estimate how the policies would affect the chance an individual would get or die from the disease. The range of answers for a particular policy (annual FOBT and flexible sigmoidoscopy) could hardly be more broad, some experts believing the screening strategy would virtually eliminate the disease (reducing mortality 95 percent), others believing it would have virtually no effect at all (reducing mortality only 5 percent) (Exhibit 2).

This example is not presented to embarrass experts; it simply indicates how the complexity of a medical problem, and the lack of good information on the basic factors needed to solve the problem, can leave us with virtually no understanding of the magnitude of the impact of a procedure on even the most important outcomes.

The difficulty of processing the numbers of probabilities important to medical reasoning is indicated by more formal studies, even for very simplified problems. Some examples follow.

A group of practicing physicians was surveyed to assess their knowl-

---

**Exhibit 2**

**Effect Of Screening: Annual Fecal Occult Blood Test And Annual Flexible Scope**

<table>
<thead>
<tr>
<th>Reduction in incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reduction in mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
</tr>
</tbody>
</table>
edge of statistical and logical skills considered necessary for interpreting the medical literature. The survey indicated the physicians had a poor understanding of commonly used terms (scoring just more than 50 on a scale of 0 to 100 in ability to define basic terms such as “false positive/negative rates,” “prevalance,” “incidence,” and “p values”) and were unfamiliar with important principles of inference or logic (answering correctly only about 25 percent of the questions about the importance of sample size, the effect of pretest probability on the predictive value of test results, and so forth).

Medical students, residents, and attending physicians at several teaching hospitals were given the following hypothetical problem: “If a test to detect a disease whose prevalence is 1/1000 has a false-positive rate of 5 percent, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person’s symptoms or signs?” Only 18 percent gave the correct answer (about a 2 percent chance). The average answer overestimated the likelihood of disease by a factor of 30.

Physicians were asked to interpret a negative computerized tomographic (CT) scan of the brain for a patient with lung cancer who was having brief reversible strokes due to reduced blood supply to the brain (transient ischemic attacks). The question presented was whether the strokes were caused by spread of the cancer or by an underlying disease of the blood vessels supplying the brain (cerebrovascular disease). A conclusion predicted by quantitative analysis indicated that the chance the symptoms were caused by spread of cancer to the brain was 98 percent. However, almost two-thirds of the physicians surveyed estimated that chance to be less than 25 percent, and only three of the forty-four physicians estimated the chance to be greater than 75 percent. An autopsy confirmed the cancer had spread to the brain.

A group of 124 physicians was given seven written case descriptions and asked to estimate the probabilities of specific diagnoses or clinical outcomes. For each question asked about the case, the physicians’ estimates varied widely, the smallest range being 80 percent (ranging from 2 percent to 82 percent). For four of the cases, the range of estimates exceeded 90 percent.

Evidence on preferences. The available information on patient preferences is even less complete. A handful of researchers in medical decision making and psychology have documented the importance and variability of patient preferences relating to a very small number of health problems. However, we have virtually no systematic understanding of how people value different types of health outcomes. For example, consider a possible policy to screen women from age forty to fifty with annual mammograms and breast examinations. To our knowledge, no one has ever asked forty-year-old women how they compare (1) reducing
their chance of dying of breast cancer that might occur in that age decade from 50 in 10,000 (without screening) to 30 in 10,000 (with screening for ten years)–a reduction of 40 percent–versus (2) a 2 percent chance each year of a false-positive test result leading to a breast biopsy, a one in a million chance of the radiation causing a new case of breast cancer, and a bill for the mammogram and breast examination ranging from about $350 to $1,700 (that is, $40 to $200 per examination for ten years, discounted). It is simply not a standard practice to present the outcomes of different choices to people, and ask their preferences.

Implications. The inescapable conclusion is that some practice standards are created without knowledge of the actual impact of the practice on health and economic outcomes and without knowledge of how people would compare the benefit and harm. This conclusion has two corollaries. First, for some current practices, decisions and standards must be based on a logic that does not use information on the magnitude of their benefit and harm, because that information is not available. Second, to the extent that patients and physicians would choose different practices if they were better informed, current practices and standards might not be appropriate.

On the first point, the reasoning used to support many clinical decisions and standards appears to be much simpler than described above. The logic appears to be that a practice will be considered appropriate if it might have benefit. For convenience, we will call this the criterion of potential benefit. This criterion is tempting for several reasons. Because it does not require estimation of the magnitude of benefit or explicit comparison of benefit and harm, it is very easy to apply. It also deals quite smoothly with the uncertainty that surrounds many practices. One can be quite uncertain about whether a practice has benefit, and, if it does, the magnitude of any benefit, and still conclude that it might have benefit. This criterion translates easily into “when in doubt, do it.”

What harm is done? The criterion of potential benefit works well for practices for which the benefit is so obvious and great and the harm (and cost) is so small that no formal trials or sophisticated analyses are needed to appreciate their roles. The rabies vaccine is a good example—a person infected by rabies will live with the vaccine and will die without the vaccine. We do not have to conduct a randomized trial to prove this or conduct a scientific survey to conclude that virtually everyone would consider the treatment worth the pain of the shots.

However, the criterion of potential benefit has at least five problems. First, it does not work for decisions or standards about practices for which the benefit does not obviously outweigh the harm. Furthermore, the criterion of potential benefit has the insidious quality of not alerting us when the benefit might not outweigh the harm, because it does not
require careful examination of the evidence and examination of the benefit and harm. It is easy to overestimate the benefit and underestimate the harm, perhaps due to optimism and wishful thinking. To the extent that we are promoting practices for which the benefit is small (or nonexistent) compared to the harm, we are doing harm.

Second, the criterion of potential benefit does not work when we must choose between competing practices. For example, a patient suffering from vascular disease in the legs can be treated by PTA or bypass. Both might cause benefit, which explains why each procedure has its advocates. However, the criterion of potential benefit provides no basis for choosing between them. To the extent that an explicit examination of outcomes of the two procedures would lead to a clear preference, and to the extent the nonpreferred practice is being used, harm is being done.

Third, the criterion of potential benefit drives up costs. If we attempt to do everything that might have benefit, and if we always resolve uncertainty in favor of potential benefit, we can end up performing and paying for practices that have little or no value.

Fourth, the criterion of potential benefit leaves us helpless when resources are limited and difficult choices must be made. The harm caused by this is inefficiency. Failure to give priority to practices that deliver the greatest health benefit for the resources can result in a net loss of benefit compared with what could potentially be gained with a more efficient allocation of resources.

The fifth problem is that use of the criterion of potential benefit can cause us to overestimate our knowledge. If the criterion of potential benefit is used, we can easily have enough knowledge to make decisions without in fact knowing very much.

**Why do we not have better evidence?** First, it should be recognized that collecting good evidence can be extremely difficult, even with the best planning. For example, it is clearly not possible to answer all clinical questions with randomized controlled trials. For some questions, the infrequency of the outcomes (requiring large sample sizes), the long natural histories of the diseases (requiring long follow-up periods), rapid changes in practices (sometimes invalidating results before completion), variations across patients, and high costs, can all conspire to make a randomized controlled trial impossible or useless. Nonetheless, many questions can be examined with randomized controlled trials, and for those that cannot, there are other designs that could yield considerably better information than is now available. For example, PTA and bypass surgery could be studied by controlled trials. Thus it is still reasonable to ask why the available evidence is not better than currently exists.

There are two basic reasons: practitioners and policymakers do not demand better evidence, and many clinical researchers do not have incentives to produce it. On the first point, practitioners who use the
criterion of potential benefit have little need for better evidence; they can set standards without ever estimating the effect of the contemplated practice on health outcomes. This in turn removes the major incentive to demand good evidence. On the second point, clinical researchers often lack the incentives to produce results useful for estimating the impact of a practice. To appreciate how this can occur, consider why there are more than three dozen clinical series on PTA but no studies that answer the important clinical question of how PTA compares with bypass surgery. When designing, conducting, and reporting research, the investigator has several goals. One goal is to learn more about PTA. But other goals might be to get a publication, get promotion or tenure in a medical school, be recognized as a “pioneer” in the use of PTA, be invited to conferences, and so forth.

How Do We Get Better Evidence?

Getting better evidence will be very difficult. Not only will it require considerable effort and money, it also will require changes in perceptions and deeply held traditions. The basic approach to getting better evidence follows from the basic reasons we do not have good evidence now—we must demand better evidence and create incentives for researchers to produce it. Some of the actions required to accomplish this are: promoting programs to examine practices; increasing coordination of research; requiring higher standards for publication; and requiring higher standards for policies.

Promote programs to examine practices. Formal examination of practices is best accomplished through programs to design practice standards and coverage policies. Specialty societies and large groups (for example, health maintenance organizations, or HMOs) can lead the effort to design practice standards. Third-party payers and purchasers can provide leadership in designing coverage policies. In all cases there should be a commitment to three things. First, the policies should be based on evidence, not group opinion, expert testimony, or “a widespread consensus.” Second, the analysis should estimate explicitly the magnitudes of the effects of different options on the important health and economic outcomes. The analysis should not be based on the mere finding that a practice might have benefit (the criterion of potential benefit). Third, the policy statement should contain the estimates of the outcomes, in addition to the conclusion (for example, the practice recommendation or coverage policy). A good strategy is to separate the estimation of outcomes from the value judgments about those outcomes. First list the health outcomes affected by the practice, then estimate the effect of the practice on each health outcome, and then make the value judgments based on the estimated effects of practice. Over the long run,
this approach will force us to examine the evidence supporting each practice, which in turn will stimulate better research. Immediate benefits are that it will help us understand the effects of what we are doing and put us in a better position to choose between practices.

**Increase coordination of research.** Health care research today is highly decentralized. By and large, each investigator conceives, designs, conducts, and reports the research as he or she chooses—subject only to availability of funding and potential for publication. Whatever this approach might gain in creativity, it loses in impact; the results frequently are not useful for clinical decisions. Great gains could be made by setting priorities for clinical research, tying research to decisions, and increasing coordination between researchers (for example, agree on common patient selection criteria, definitions, follow-up times, and reporting methods).

**Higher standards for publication.** Standards for publication in the handful of best journals are already tight (although they can be tightened further). However, a far more pressing problem is to achieve tighter standards for the second-line journals. There are two main goals. One is well recognized: to ensure that any study published is well designed, conducted, analyzed, and reported. The other goal, less well achieved, is to increase the usefulness of published studies for making clinical decisions. This requires not only that the study’s designs be internally sound, but also that it be motivated by the practical questions and that the information it provides can be tightly fitted with the results of other studies. Somehow, if a study’s results are to be useful, they must be integrated with existing knowledge and other research. This integration should be addressed before the study is designed, not after. If publication of nonuseful reports were made difficult, researchers would move quickly to improve the usefulness of their work. Journal editors can lead this task.

**Higher standards for standards.** Immense care can go into designing a health technology and collecting evidence about its effectiveness; yet there are almost no standards for what individuals and organizations can say about its appropriate use. The result is a strong tendency to take the easy route, which does not require careful evaluation of the evidence or explicit estimation of the consequences of different options. To force a more careful evaluation of the existing evidence and increase demand for better evidence, we must tighten the “standards for standards.”

**Leadership**

Who decides if the problems described in this article are real and serious? Who decides if these or other suggestions are reasonable? Who has the authority or the resources for implementing whatever is agreed
on? Who is accountable fifteen years from now if no progress has been made? A basic fact of the American health care system is that there are no good answers to these questions. This is because our system does not have a leader; it has dozens—each with different but overlapping domains and interests. Furthermore, there is no mechanism for both developing and executing a common vision.

To improve the quality of evidence, we must either change the system to create such a mechanism. Or the current leaders of the system must coalesce around this issue. However it is accomplished, we must all agree on the principles that should guide our health care system, on any discrepancies between the principles and our actions, on the steps we need to correct those discrepancies, and on the sacrifices and compromises we must make to put our principles into operation. Until this is achieved, our progress toward improving the quality of evidence, the quality of standards, and the quality of the health care people actually receive, will be mixed, and slow.

NOTES

1. The term “standard” will be used in a general sense to describe a widely accepted statement that unambiguously defines the appropriate use of a particular practice. Related terms are “rules,” “policies,” and “guidelines.” The last two terms indicate greater room for differences of opinion and variations in a practice. The term “practice” will be used in a general sense to designate any intervention designed to maintain or improve a person’s health. It includes prevention, screening, diagnostic, treatment, and supportive activities. A synonym is “health technology.”


