VARIATIONS IN
PHYSICIAN PRACTICE:
THE ROLE OF
UNCERTAINTY
by David M. Eddy

Prologue: Nothing more underscores the reality that medical care represents a melding of art and science than the wide variation in the use and cost of that care. One of the factors that leads to practice variations is uncertainty in the minds of doctors about how to approach a particular medical problem. David Eddy, who holds a medical degree from the University of Virginia and a doctor of philosophy degree in applied mathematics from Stanford University, has made a professional career of reducing uncertainty in medical practice and helping doctors find their way to the most efficacious treatments for patients. Eddy operates at the intersection of math and medicine, applying probability theory to the uncertainty of approaches to care. Although his pursuits have not always been appreciated by the profession—many of his conclusions cut against the grain of conventional medical judgment—Eddy has been awarded the Lanchester Prize by the Operations Research Society of America, the most prestigious honor that this organization of applied mathematicians bestows. Eddy was recognized for a book he authored entitled, Screening for Cancer: Theory, Analysis and Design. The essence of the book was a study undertaken by the American Cancer Society to determine how often a doctor should give women a pap smear. For forty years, the society has promoted the annual pap smear as a cancer detection treatment. Factoring in all of the relevant data and subjecting it to his mathematical microscope, Eddy determined that the most sensible frequency for administering this test is every three years. The cancer society changed its recommendations accordingly, provoking a storm of controversy within some medical quarters. The American College of Obstetrics and Gynecology took particular exception to the recommended three-year frequency, but the cancer society did not waver from its new standard. Eddy is a member of the Institute of Medicine of the National Academy of Sciences and is a member of the World Health Organization Expert Advisory Panel on Cancer. Eddy is now developing a third-party payer process of setting reimbursement policies for medical procedures.
Why do physicians vary so much in the way they practice medicine? At first view, there should be no problem. There are diseases—neatly named and categorized by textbooks, journal articles, and medical specialty societies. There are various procedures physicians can use to diagnose and treat these diseases. It should be possible to determine the value of any particular procedure by applying it to patients who have a disease and observing the outcome. And the rest should be easy—if the outcome is good, the procedure should be used for patients with that disease; if the outcome is bad, it should not. Some variation in practice patterns can be expected due to differences in the incidence of various diseases, patients’ preferences, and the available resources, but these variations should be small and explainable.

The problem of course is that nothing is this simple. Uncertainty, biases, errors, and differences of opinions, motives, and values weaken every link in the chain that connects a patient’s actual condition to the selection of a diagnostic test or treatment. This paper describes some of the factors that cause decisions about the use of medical procedures to be so difficult, and that contribute to the alarming variations we observe in actual practice. It examines the components of the decision problem a physician faces, and the psychology of medical reasoning, focusing in particular on the role of uncertainty. Finally, it suggests some actions to reduce uncertainty and encourage consistency of good medical practice.

Uncertainty creeps into medical practice through every pore. Whether a physician is defining a disease, making a diagnosis, selecting a procedure, observing outcomes, assessing probabilities, assigning preferences, or putting it all together, he is walking on very slippery terrain. It is difficult for nonphysicians, and for many physicians, to appreciate how complex these tasks are, how poorly we understand them, and how easy it is for honest people to come to different conclusions.

Defining A Disease

If one looks at patients who are obviously ill, it is fairly easy to identify the physical and chemical disorders that characterize that illness. On the other hand, a large part of medicine is practiced on people who do not have obvious illnesses, but rather have signs, symptoms, or findings that may or may not represent an illness that should be treated. Three closely related problems make it difficult to determine whether or not a patient actually has a disease that needs to be diagnosed or treated.

One problem is that the dividing line between “normal” and “abnormal” is not nearly as sharp as a cursory reading of a textbook would suggest. First, the clues on which we base the diagnosis of many diseases can be

The preparation of this paper was supported by The John A. Hartford Foundation.
very difficult to see, with frequent errors in both directions (missing an existing disease and “finding” a nondisease). Second, even if the diagnosis were correct and a disease were acknowledged to be present, the “disease” might not actually cause the patient any harm. Dysplasia of the cervix is a good example of both problems. It is an abnormal finding in the sense that most women do not have it, and it is associated with the development of cancer of the cervix. On the other hand, it is notoriously difficult to diagnose with certainty because dysplastic cells are only slightly different in appearance than normal cells (see example to follow under “Making A Diagnosis”), and in the majority of cases it disappears spontaneously (assuming it was there in the first place). Obesity, hyperplasia of the tonsils, fibrocystic disease of the breast, and dozens of other conditions pose similar dilemmas.

A second problem is that many “diseases,” at least at the time they are diagnosed, do not by themselves cause pain, suffering, disability, or threat to life. They are considered diseases only because they increase the probability that something else that is truly bad will happen in the future. This raises two more sources of uncertainty. (1) If a condition presages a bad outcome, one must judge the probabilities. Most conditions of this type do not always cause a “real” disease, and the “real” disease can usually occur without the condition. In situ lobular carcinoma of the breast presages a future invasive breast cancer less than 50 percent of the time, and the great majority of invasive breast cancers occur without a history of in situ lobular carcinoma. (2) Just because a condition can precede a “real” disease and can indicate a higher probability that the disease will develop, does not necessarily mean that it causes the disease, or that treating the condition will prevent it from occurring. Ocular hypertension and glaucoma are good examples; loss of visual field and blindness appear to occur whether or not the ocular pressure is lowered.

The difficulty of defining a disease is compounded by the fact that many of the signs, symptoms, findings, and conditions that might suggest a disease are extremely common. If a breast biopsy were performed on a random sample of senior citizens, fully 90 percent of them could have fibrocystic disease. If obesity is a disease, the average American is diseased, By the time they reach seventy, about two-thirds of women have had their uteruses removed. Because the average blood pressure increases with age, some physicians feel a need to relabel “hypertension” to keep the majority of older people from having this disease.

And the ambiguities grow worse as medical technology expands. More and more diseases are being defined by an abnormal result on some test, leaving uncertainty about its real meaning to a patient and the appropriate treatment Silent gallstones were silent until the oral cholecystogram was introduced, dysplasia of the cervix did not exist before the Pap smear, and many people’s coronary artery disease showed up only on a
treadmill test. Finding “diseases” early may be worthwhile, but it is difficult to know what else is being scooped up in the net.

Given these uncertainties about what constitutes a disease, it should not be surprising that there are debates about the definitions of many diseases, and when there is agreement about a definition, it is often blatantly and admittedly arbitrary. A quick review of the literature reveals multiple definitions of glaucoma, diabetes, fibrocystic disease of the breast, coronary artery disease, myocardial infarction, stroke, and dozens of other conditions. Morbid obesity is defined as 100 percent above the ideal weight. But what is “ideal,” and why 100 percent? The lesson is that for many conditions a clinician faces, there is no clear definition of disease that provides an unequivocal guide to action, and there is wide room for differences of opinion and variations in practice.

Making A Diagnosis

Suppose everyone agreed that a particular collection of signs, symptoms, and test results constituted an unequivocal definition of a disease. Would this eliminate the uncertainty? Unfortunately, even when sharp criteria are created, physicians vary widely in their application of these criteria—in their ability to ask about symptoms, observe signs, interpret test results, and record the answers. The literature on “observer variation” has been growing for a long time. To cite some of the classics:

Cyanosis, or blueing of the face and fingers, is considered a sign of low oxygen content in the blood. One investigator compared the abilities of twenty-two doctors to note cyanosis in twenty patients, the true diagnosis of cyanosis being confirmed by oximeter under controlled conditions. Only 53 percent of the physicians were definite in diagnosing cyanosis in subjects with extremely low oxygen content. And 26 percent of the physicians said cyanosis existed in subjects with normal oxygen content.

Perhaps the error rates are less severe if the physician can study “hard” evidence like x-rays, electrocardiograms, or laboratory procedures:

A set of 1,807 photofluorograms containing thirty “positive” and 1,760 “negative” films (as defined by unanimous agreement of seven experts), were read independently by ten physicians. As many as 32 percent of the positive films were missed, and 2 percent of the negative films (thirty-five films) were incorrectly called positive. When individual readers read the same films on two separate occasions, they disagreed with themselves about 20 percent of the time.

A group of experts compiled 100 electrocardiogram tracings, fifty of which showed myocardial infections, twenty-five of which were normal, and twenty-five of which showed some other abnormality (according
to the experts). These EKGs were then given to ten other cardiologists to test their diagnostic abilities. The proportion of EKGs judged by the ten cardiologists to show infarcts varied by a factor of two. If you had an infarct and went to physician A, there would be a 28 percent chance the physician would have missed it. If you did not have an infarct and went to physician B, there would be a 26 percent chance that physician would have said you had one.

How much confidence can we have in taking a person’s medical history? Four physicians interviewed 993 coal miners about several common symptoms, including cough, sputum, shortness of breath, and pain. After each physician completed all his interviews, he was asked to record the proportion of miners who reported each symptom (for example, to state the proportion of miners who answered yes to the question, “Do you have a cough?”). The proportion of miners reported to have various symptoms varied from 23 percent to 40 percent for cough, 13 percent to 42 percent for sputum, 10 percent to 18 percent for shortness of breath, and 6 percent to 17 percent for pain.

Perhaps the hard eye of the microscope can yield definitive answers. Thirteen pathologists were asked to read 1,001 specimens obtained from biopsies of the cervix, and then to repeat the readings at a later time. On average, each pathologist agreed with himself only 89 percent of the time (intraobserver agreement), and with a panel of “senior” pathologists only 87 percent of the time (interobserver agreement). Looking only at the patients who actually had cervical pathology, the intraobserver agreement was only 68 percent and the interobserver agreement was only 51 percent. The pathologists were best at reading more advanced disease and normal tissue, but were labeled “unsatisfactory” in their ability to read the precancerous and preinvasive stages.

Similar studies have been reported for the presence of clubbing of the fingers, anemia, psychiatric disease, and many other signs, symptoms, and procedures. Even if there were no uncertainty about what constitutes a disease and how to define it, there would still be considerable uncertainty about whether or not a patient has the signs, symptoms, and findings needed to fit the definition.

**Selecting A Procedure**

The task of selecting a procedure is no less difficult. There are two main issues.

First, for any patient condition there are dozens of procedures that can be ordered, in any combination, at any time. The list of procedures that might be included in a workup of chest pain or hypertension would take more than a page, spanning the spectrum from simply asking questions,
to blood studies, to x-rays. Even for highly specific diagnostic problems, there can be a large choice of procedures. For example, if a woman presents with a breast mass and her physician wants to know its approximate size and architecture, the physician might contemplate an imaging procedure. The choice could include mammography, ultrasonography, thermography, diaphanography, computed tomography, lymphography, Mammoscan, and nuclear magnetic resonance imaging. A physician who chose mammography would still have to decide between xeromammography and film mammography, with several brands being available for the latter. There are about a dozen procedures that apply the principles of thermography. And why should a diagnostic workup be limited to one test? Why not follow a negative mammogram with a computed tomogram (or vice versa)? For the detection of colorectal cancer, a physician can choose any combination of fecal occult blood tests (and there are more than a dozen brands), digital examination, rigid sigmoidoscopy, flexible 30 cm sigmoidoscopy, flexible 60 cm sigmoidoscopy, barium enema (either plain or air contrast), and colonoscopy. These choices are not trivial. Most procedures have different mechanisms of action and a long list of pros and cons. Different brands of fecal occult blood tests have very different sensitivities and specificities, and film mammography and xeromammography differ in their radiation exposure by a factor of about four. These procedures are for relatively well-defined diseases; imagine the problems of selecting procedures to evaluate symptoms like fatigue, headache, or fever that can have about a dozen causes.

Second, adding to the uncertainties of defining and choosing a procedure is the fact that the value of any particular procedure depends on who performs it, on whom it is performed, and the circumstances of performance. The potential for variability in the people who perform procedures can be appreciated by considering one of the simplest procedures, the Pap smear. A gynecologist reviewed the technique used by sixty of his colleagues to take a Pap smear, and found that only fifteen of them performed the test properly. With this amount of slippage in such a simple test, one can only imagine the variation in quality that occurs with a more complicated procedure like coronary artery bypass surgery. With respect to who receives the procedure, the outcome of a test will depend on the probability the patient has the disease in question, on the physical condition of the patient for example, young breasts absorb x-rays differently than older breasts, and on the patient’s psychological condition (some people can tolerate passing a colonoscope all the way to the cecum, while others cannot). Finally, the circumstances under which a procedure is performed can have a dramatic effect on its value. Blood pressures go up for insurance examinations. Ocular pressures fluctuate by several millimeters of mercury every day. An IQ test can be a joy for a person with a good night’s sleep, and a tragedy for a person with a head cold. The
message is that a “procedure” is not a procedure. Each procedure has many faces, and many factors influence the quality and consequences of its use.

Observing Outcomes

In theory, much of the uncertainty just described could be managed if it were possible to conduct enough experiments under enough conditions, and observe the outcomes. Unfortunately, measuring the outcomes of medical procedures is one of the most difficult problems we face. The goal is to predict how the use of a procedure in a particular case will affect that patient’s health and welfare. Standing in the way are at least a half dozen major obstacles. The central problem is that there is a natural variation in the way people respond to a medical procedure. Take ‘two people who, to the best of our ability to define such things, are identical in all important respects, submit them to the same operative procedure, and one will die on the operating table while the other will not. Because of this natural variation, we can only talk about the probabilities of various outcomes—the probability that a diagnostic test will be positive if the disease is present (sensitivity), the probability that a test will be negative if the disease is absent (specificity), the probability that a treatment will yield a certain result, and so forth.

One consequence of this natural variation is that to study the outcomes of any procedure it is necessary to conduct the procedure on many different people who are thought to represent the particular patients we want to know about, and then average the results. This in turn raises additional problems. First, many of the diseases are fairly rare, and it is necessary to average over many people to get a sample large enough to yield reliable results. This usually requires using many physicians, drawing patients from many settings, and performing the experiments at different times. Each of these elements introduces additional variation. Some diseases are so rare that, in order to conduct the ideal clinical trials, it would be necessary to collect tens of thousands, if not hundreds of thousands, of participants. A good example concerns the frequency of the Pap smear. One might wonder why the merits of a three-year versus one-year frequency cannot be settled by a randomized controlled trial. Because of the low frequency of cervical cancer, and the small difference in outcomes expected or the two frequencies, almost one million women would be required for such a study.

An additional problem is that most procedures have multiple outcomes and it is not sufficient to examine just one of them. For example, a coronary artery bypass may change the life expectancy of a sixty-year-old man with triple vessel disease, but it will also change his joy of life for several weeks after the operation, the degree and severity of his chest
pain, his ability to work and make love, his relationship with his son, the physical appearance of his chest, and his pocketbook. Pain, disability, anxiety, family relations, and any number of other outcomes are all important consequences of a procedure that deserve consideration. But the list is too long for practical experiments and many of the items on it are invisible or not measurable at all. We either lack suitable units (for example, for anxiety or pain), or the units exist but no experiments are fine enough to detect a change (for example, the increased incidence in breast cancer due to radiation from mammography).

Beyond this, many of the outcomes needed to evaluate a medical procedure take years to observe. There is no way to measure the ten-year survival of a patient following a porto-caval shunt without waiting ten years. To pursue the example of the Pap smear, the long duration of the preinvasive stages of the disease means that if the study with one million women were initiated, it would have to be continued for more than two decades to learn the results.

Finally, even when the best trials are conducted, we still might not get an answer. Consider the value of mammography in women under fifty, and consider just one outcome—the effect on breast cancer mortality. Ignore for the time being the radiation hazard, false-positive test results, inconvenience, financial costs, and other issues. This is one of the best-studied problems in cancer prevention, benefiting from the largest (60,000 women) and longest (more than fifteen years) completed randomized controlled trial, and an even larger uncontrolled study involving 270,000 women screened for five years in twenty-nine centers around the country. Yet we still do not know the value of mammography in women under fifty. The first study showed a slight reduction in mortality, but it was not statistically significant. The larger study suggested that mammography has improved since the first study, and that it is now almost as good in younger women as in older women, but the study was not controlled and we do not know if “almost” is good enough. Even for women over fifty, where the first study showed a statistically significant reduction in breast cancer mortality (of about 40 percent at ten years), there is enough uncertainty about the results that no fewer than four additional trials have been initiated to confirm the results. These trials are still in progress.

Unable to turn to a definitive body of clinical and epidemiological research, a clinician or research scientist who wants to know the value of a procedure is left with a mixture of randomized controlled trials, nonrandomized trials, uncontrolled trials, and clinical observations. The evidence from different sources can easily go in different directions, and it is virtually impossible for anyone to sort things out in his or her head. Unfortunately, the individual physician may be most impressed by observations made in his or her individual practice. This source of evidence is notoriously vulnerable to bias and error. What a physician sees and
remembers is biased by the types of patients who come in; by the decisions of the patients to accept a treatment and return for follow-up; by a natural desire to see good things; and by a whole series of emotions that charge one’s memory. On top of these biases, the observations are vulnerable to large statistical errors because of the small number of patients a physician sees in a personal practice.

The difficulty of measuring outcomes has three important implications: We are uncertain about the precise consequences of using a particular procedure for a particular patient. We cannot, over the short term at least, resolve this uncertainty. And whatever a physician chooses to do cannot be proved right or wrong.

Assessing Preferences

Now assume that a physician can know the outcomes of recommending a particular procedure for a particular patient. Is it possible to declare whether those outcomes are good or bad? Unfortunately, no. The basic problem is that any procedure has multiple outcomes, some good and some bad. The expected reduction in chest pain that some people will get from coronary artery bypass surgery is accompanied by a splitting of the chest, a chance of an operative mortality, days in the hospital, pain, anxiety, and financial expense. Because the outcomes are multiple and move in different directions, tradeoffs have to be made. And making tradeoffs involves values.

Just as there is a natural variation in how each of us responds to a medical procedure, there is a variation in how we value different outcomes. The fact that General Motors alone produces more than fifty distinct models of automobiles, not to mention dozens of options for each model, demonstrates how tastes about even a single item can vary. Imagine the variation in how different people value pain, disability, operative mortality, life expectancy, a day in a hospital, and who is going to feed the dogs.

In fact, for the outcomes of medical procedures, variations in the values of different people can be huge. Consider a single outcome of a fairly simple procedure—the scar from a breast biopsy. One of the ingredients to a physician’s decision to recommend a biopsy for a woman with a breast mass is the physician’s assessment of how the woman values the cosmetic affects of the surgery. How important is it to her not to have a small scar on her breast? While it is difficult to know precisely, one can pose questions such as the following to women.

Pretend that you have just had a breast biopsy. You have already received the results of the biopsy and know that you do not have cancer. There is no more medical information to be obtained from further studies. However, following the biopsy, you have on the upper outer quadrant (at about 3 o’clock) of your left breast a small one-inch
scar that is slightly indented from the removal of a piece of tissue about the size of a pecan. I am a wizard and I can snap my fingers and make that scar disappear without a trace. I cannot erase the memory of your hospitalization, any anxiety you had prior to surgery, or any of the other events surrounding your biopsy, but if I snap my fingers, your scar will disappear. How much will you pay me to snap my fingers?

When about twenty women were asked this question in an informal setting, the answers ranged from less than $100 to $10,000. In addition to the wide variation in the answers, it is pertinent that husbands typically gave lower numbers than their wives, and physicians gave the lowest numbers of all.

To the inherent variation in values individual patients place on different outcomes must be added two additional sources of uncertainty and variation in assessing values. First, because decisions about procedures are typically made by physicians on behalf of their patients, the physicians must infer their patients’ values, and keep them distinct from their own personal preferences. This raises the second problem, communication. It is difficult enough to assess one’s own values about the outcomes of a complicated decision (think about switching jobs); consider having someone else try to learn your thoughts and do it for you. The room for error in communication can be appreciated by returning to the experiment in which four physicians asked 993 coal miners about cough, shortness of breath, pain, and sputum. The variation in their reports of responses to a simple question like, “Do you have a cough?” was large; imagine a question like, “How do you feel about operative mortality?”

Putting It All Together

The final decision about how to manage a patient requires synthesizing all the information about a disease, the patient, signs and symptoms, the effectiveness of dozens of tests and treatments; outcomes, and values. All of this must be done without knowing precisely what the patient has, with uncertainty about signs and symptoms, with imperfect knowledge of the sensitivity and specificity of tests, with no training in manipulating probabilities, with incomplete and biased information about outcomes, and with no language for communicating or assessing values. If each piece of this puzzle is difficult, it is even more difficult for anyone to synthesize all the information and be certain of the answer. It would be an extremely hard task for a research team; there is no hope that it could occur with any precision in the head of a busy clinician. Hence the wide variability in the estimates physicians place on the values of procedures.

Two final examples document how difficult it is to combine information from many sources to estimate the value of a particular procedure. The fecal occult blood test can be used to detect blood in the stool of
Exhibit 1
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>

asymptomatic people for the early detection of colorectal cancer. Flexible sigmoidoscopy (60 cm) can detect precancerous adenomas and cancers. At a recent meeting of experts in colorectal cancer detection, all of whom were very familiar with fecal occult blood testing (and most of whom had participated in two previous meetings on cancer detection in the previous four years), the attendees were asked the following question: “What is the overall reduction in colorectal cancer incidence and mortality that could be expected if men and women over the age of fifty were tested with fecal occult blood tests and 60 cm flexible sigmoidoscopy every year?” The answer to this question is obviously central to any estimate of the value of fecal occult blood testing, and it is pertinent that the experts were unanimous in their belief that the fecal occult blood test was valuable and should indeed be recommended annually to men and women over fifty. The answers were distributed as shown in Exhibit 1. It is tempting to say that nonexperts, or people who did not share the belief that the fecal occult blood test is valuable, would have shown a wider variation, but the variation expressed by this group could hardly be any wider. As startling as the degree of variation in the estimates, is that the attendees were surprised by the results; they had never communicated this number to each other, and had no idea they had such differences of opinion.

The second example is a classic. A survey of 1,000 eleven-year-old schoolchildren in New York City found that 65 percent had undergone tonsillectomy. The remaining children were sent for examinations to a group of physicians and 45 percent were selected for tonsillectomy. Those rejected were examined by another group of physicians and 46 percent were selected for surgery. When the remaining children were examined again by another group of physicians, a similar percent were recommended
for tonsillectomy, leaving only sixty-five students. At that point, the study was halted for lack of physicians.

Consequences

The view of anyone who wants a close look at the consequences of different medical procedures is, at best, smoky. Some procedures may present a clear picture, and their value, or lack of it, may be obvious; putting a finger on a bleeding carotid artery is an extreme example. But for many, if not most medical procedures, we can only see shadows and gross movements. We usually know the direction in which various outcome measures can move when a medical activity is undertaken, but we typically do not know the probabilities they will move in those directions, or how far they will move. We certainly do not know how a particular individual will respond. Words like “rare,” “common,” and “a lot” must be used instead of “One out of 1,000,” or “seven on a scale of one to ten.”

There is also a strong tendency to oversimplify. One of the easiest ways to fit a large problem in our minds is to lop off huge parts of it. In medical decisions, one option is to focus on length of life and discount inconvenience, pain, disability, short-term risks, and financial costs. A physician can also draw on a number of simplifying heuristics. Anyone uncomfortable dealing with probabilities can use the heuristic, “If there is any chance of (the disease), the (procedure) should be performed.” If one cannot estimate the number of people to be saved, one can use the heuristic, “If but one patient is saved, the effort is worthwhile.” If one cannot contemplate alternative uses of resources that might deliver a greater benefit to a population, there is the heuristic, “Costs should not be considered in decisions about individual patients.” There is a general purpose heuristic, “When in doubt, do it.” Or as one investigator wrote, “An error of commission is to be preferred to an error of omission.”

Unfortunately, a large number of incentives encourage simplifications that lead to overutilization. It is time-consuming, mentally taxing, and often threatening to colleagues for a physician to undertake a deep analysis of a confusing clinical problem. A physician is less likely to be sued for doing too much than too little. Most physicians, incomes go up if they do more, and go down if they do less. Hospitals get to fill more beds and bill for more procedures, laboratories collect more money for services, and companies sell more drugs, devices, and instruments. The more that is done, the more the providers win. The losers are patients, consumers, and taxpayers—anyone who has to undergo a valueless procedure or pay the bill.

In the end, given all the uncertainties, incentives, and heuristics, a physician will have to do what is comfortable. If it is admitted that the uncertainty surrounding the use of a procedure is great, and that there is no
way to identify for certain what is best, or to prove that any particular action is right or wrong, the safest and most comfortable position is to do what others are doing. The applicable maxim is “safety in numbers.” A physician who follows the practices of his or her colleagues is safe from criticism, free from having to explain his or her actions, and defended by the concurrence of colleagues.

This tendency to follow the pack is the most important single explanation of regional variations in medical practice. If uncertainty caused individual physicians to practice at random, or to follow their personal interpretations and values, without any attempts to match the actions of their neighbors, the variations in practice patterns would average out, and no significant differences would be observed at the regional level. Differences between regions are observed because individual physicians tend to follow what is considered standard and accepted in the community. A community standard evolves from statements published in national journals and textbooks, from the opinions of established physicians, and from new ideas brought to the community by new physicians. The community standards themselves exist because enough is known to enable the leaders of a community to develop opinions which, when followed by their colleagues, become community standards. The differences between community standards exist because not enough is known to establish which opinion is correct. We call the community standards for a particular practice clinical policies, and anyone who makes an unambiguous recommendation about a medical practice is a policymaker.

**What Harm Is Done?**

First, it should be clear that some variation in practice is appropriate. The differences in patients’ risks, signs and symptoms, responses to treatment, and values are real. Differences in physicians' talents and the available facilities are also real. If physicians were able to tailor their practices to take these individual differences into account, variations would be both inevitable and desirable. The problem is that uncertainty so clouds every aspect of this problem that many of the appropriate variations cannot occur, and many of the variations we see are not motivated by logic or a deep understanding of the issues.

There is no doubt that uncertainty about the consequences of different medical activities can harm both the quality and cost of medical practice. It is also true, however, that most of the simplifications and heuristics point in one direction, toward overutilization. When this happens the price is paid in terms of inconvenience, pain, distress, days in the hospital, unnecessary risks, and money.
Conclusions

Many of the problems described in this paper are insurmountable. There is no way to shorten the time needed to observe ten-year survival rates, and there is no way to increase the frequency of rare diseases, reduce the number of outcomes that are important to a patient, or decrease natural variations in response to treatment. Nor do we want to suppress the differences that exist in patients’ preferences.

However, while we can not eliminate uncertainty, we can decrease the amount of it and develop strategies to minimize its damage. In fact, the profession and society have not begun to exploit the available techniques for reducing uncertainty and maximizing expected outcomes. The evaluation of medical practices and the development of clinical policies deserve much more attention and a higher priority than they currently get. Wennberg and others have described how data bases and other techniques can improve the available information. The next task is to improve our ability to process the information we get. This calls for several actions, all designed to develop a tradition that insists on the collection and evaluation of information to understand and describe the consequences of medical practices.

First, physicians can do more to admit the existence of uncertainty, both to themselves and to their patients. While this will undoubtedly be unsettling, it is honest, and it opens the way for a more intensive search for ways to reduce uncertainty.

Second, people who want to promote policies regarding the use of medical procedures can learn the necessary languages. Over the past few hundred years languages have been developed for collecting and interpreting evidence (statistics), dealing with uncertainty (probability theory), synthesizing evidence and estimating outcomes (mathematics), and making decisions (economics and decision theory.) These languages are not currently learned by most clinical policymakers; they should be.

Third, physicians who follow existing policies can examine more carefully the supporting evidence and logic. The mere fact that a policy is established and accepted does not make it correct.

Fourth, to encourage and assist the two previous actions, any policy statement, whether it be made by an individual physician at a hospital conference, or a third-party payer considering reimbursement, should be accompanied by (1) a list of medical and economic outcomes that were considered in making the policy, (2) the policymaker’s estimates of what can be expected to happen with respect to each of the listed outcomes if the policy is followed, and (3) the supporting evidence for those estimates. Any policymaker unable to supply that information should not be making policies.

Fifth, editors and reviewers of journals can encourage publication of
good papers that synthesize existing information, estimate the outcomes of different policies, and present the rationales for different actions. Such work, while not traditional, is both difficult and important.

Sixth, editors and reviewers can require that any author who recommends a policy supply the information listed in action 4. No good journal today will report the results of an experiment without a description of the design and methods; it is no less important to describe the reasoning behind a policy statement.

Seventh, the government can support far more evaluation research to analyze medical practices. The National Institutes of Health spend more than $5 billion each year to learn more about diseases and develop tests and procedures. The budget of the major federal unit charged with determining how medical procedures should be used, the National Center for Health Care Technology (NCHCT), was $4 million, less than one thousandth as large. Even that was considered too much, and the budget was recently cut to zero. Not only should the NCHCT be revived, it should be expanded by a factor of ten to 100. Research to evaluate medical practices is like the windows in a car; without them there is little way to know where you are going.

Finally, patients can push the process by asking questions. If informed of an operative mortality rate, they can ask, “What percent?” If told about the discomfort of a particular procedure, they can ask how it compares to having a tooth pulled under novocaine, or some other event they can identify with. If a procedure is recommended, a patient can ask why, what might be found, with what probability, what difference will it make, and so forth. Many physicians will be uneasy, and some even angry, when asked questions of this type because they may not know the answers. But there are few things better than asking questions to force research to get the answers.

I believe these actions should be taken. Some of the uncertainty and the resulting variations in practice patterns that exist are unavoidable, but much of the uncertainty can be managed far better than is done now. The problems that exist today are not the fault of any individuals; the fault lies with the profession and society as a whole for not developing the traditions and methods needed to assess medical practices. Today the problem is bad; five years from now, if not improved, it will be a tragedy.
NOTES