International Journal for Quality in Health Care 2000; **Volume 12, Number 4**: pp. 281–295

Defining and measuring quality of care: a perspective from US researchers

ROBERT H. BROOK[1,2], ELIZABETH A. McGLYNN[1] AND PAUL G. SHEKELLE[1]

[1] RAND, Santa Monica and [2] UCLA Center for the Health Sciences and the Greater Los Angeles Veterans Administration Health Care System, Los Angeles, CA, USA

Abstract

The modern quality field in medicine is about one-third of a century old. The purpose of this paper is to summarize what we know about quality of care and indicate what we can do to improve quality of care in the next century. We assert that quality can be measured, that quality of care varies enormously, that improving quality of care is difficult, that financial incentives directed at the health system level have little effect on quality, and that we lack a publicly available tool kit to assess quality.

To improve quality of care we will need adequate data and that will require patients to provide information about what happened to them and to allow people to abstract their medical records. It also will require that physicians provide patient information when asked. We also need a strategy to measure quality and then report the results and we need to place in the public domain tool kits that can be used by physicians, administrators, and patient groups to assess and improve quality. Each country should have a national quality report, based on standardized comprehensive and scientifically valid measures, which describes the country's progress in improving quality of care. We can act now.

For the 70–100 procedures that dominate what physicians do, we should have a computer-based, prospective system to ensure that physicians ask patients the questions required to decide whether to do the procedure. The patient should verify the responses. Answers from patients should be combined with test results and other information obtained from the patient's physician to produce an assessment of the procedure's appropriateness and necessity.

Advanced tools to assess quality, based on data from the patient and medical records, are also currently being developed. These tools could be used to comprehensively assess the quality of primary care across multiple conditions at the country, regional, and medical group level.

Keywords: health care, health services evaluation, quality assessment, quality improvement

As this century closes, it is fitting to reflect on what we know about measuring quality of care. This essay is a personal one and reflects mostly research that we have performed at RAND. After doing that, we will consider what needs to be done in the next quarter of a century.

What we have learned about quality of care

In the last 30 years, research has demonstrated that quality can be measured [1], that quality varies enormously [2], that where you go for care affects its quality far more than who you are [3], that improving quality of care, while possible, is difficult and painful [4] and, in general, has not been successfully accomplished. Changing the way one pays for care or one's system of care will affect the amount of care rendered, but its effect on the distribution of quality care is likely to be clinically unimportant [5] Finally, there is no tool kit in the public domain that can be used in non-research settings to comprehensively assess quality of care. (For a series of quality factoids, see Appendix.)

This last deficiency stems primarily from the lack of a government policy – in any country in the world – to support development of a set of quality assessment tools that are routinely updated, user-friendly, and available to any interested party. No country in the world requires the production of a yearly national report on the level of quality delivered in its health system, although all countries in the world produce multiple financial reports. A visitor from Mars might conclude that the purpose of the US health system is to spend money; and that most of health policy is about who gets the money – doctors, lawyers, or administrators.

Progress on improving quality of care can be illustrated by a story. A few years ago the first author attended a lecture

Address reprint requests to R. H. Brook, The RAND Corporation, Health Program, 1700 Main Street, PO Box 2138, Santa Monica, CA 90407–2138, USA. E-mail: robert_brook@rand.org

given by a clinical investigator who had spent his long professional life trying to find a cure for colon cancer. The title of his lecture was 'Results of Controlled Clinical Trials to Improve Survival in Patients with Colon Cancer'. This author came to the lecture fully expecting to be told that the lecturer's life had been a success and that substantial progress had been made in advancing the state of the art of treating patients with colon cancer. Instead, he began his lecture by saying, 'I do not believe my life has been a waste, because I have systematically studied, in the most rigorous way, one proposed innovation after another for the treatment of colon cancer. I have proven beyond the shadow of a doubt that none of these interventions is worth applying to patients, and many of them may actually be harmful. What I have contributed to the field is not new knowledge about what should be done; instead, I have rid the world of a great number of hypotheses about what works by showing that they do not.'

Research in quality of care assessment performed in the last third of this century should be viewed in the same positive way. What we have done is to show world leaders that quality of care varies remarkably, that probably millions of people in developed countries have shortened life expectancies because of poor quality, that neither government nor private-sector policies to improve quality have succeeded; and that efforts to improve quality have not kept up with the scientific advances we have made in medicine. Although the likelihood that a person will benefit from medical care is better now than it was a third of a century ago, largely as a result of investment in basic science and clinical research, there is no evidence that we are better today at applying what we know than we were 30 years ago. Indeed, we may be worse because the complexity of medicine has increased so greatly.

How can we change this situation? First, we need a coordinated strategy to produce, and place in the public domain, tool kits that can be used by physicians, administrators, and patient advocacy groups to assess, improve, and alter medical care quality. This will require sustained government funding because the science upon which quality measurement is based is constantly changing. Second, we need to shift some of the resources now spent on advancing the science of medicine to re-engineering our health system to improve its efficiency and effectiveness. In many instances, we would do more to improve the health of the population by using better what we know than by learning new things. Third, each country should have a national report that describes its progress in improving the quality of care. The reports should be based on standardized, comprehensive, and scientifically valid quality of care measures. The reports should highlight implications of the variations in quality that now exist, and evaluate whether we are making progress in our efforts to apply the science that has been developed, mostly with public dollars, in an efficient and effective way to improve people's health.

Defining quality care

One of the most important contributions of research in the quality field has been its attempt to define what is meant by quality care. All definitions seem to contain two components that are important to people [1]. The first component is providing care of high technical quality. By high technical quality care we mean that the patient receives only the procedures, tests, or services for which the desired health outcomes exceed the health risks by a sufficiently wide margin; and that each of these procedures or services is performed in a technically excellent manner. The second component of quality of care is that all patients wish to be treated in a humane and culturally appropriate manner and be invited to participate fully in deciding about their therapy.

Individuals' value systems and their conditions shape their choices about which component is most important. A patient with a broken leg might place the highest value on the technical aspect of care. A patient who has a chronic or an acute self-limiting condition might value the art of care over technical quality. Other preferences in the art domain, such as using local community versus distant tertiary referral centers, may dominate patient choice. For instance, a patient with cancer who might benefit from going to a referral center that has more experience in treating the tumor may choose to go to her local hospital where she knows and trusts the health professionals even if this means assuming an increased risk of a poorer outcome, including death. What we must do is give people enough information to make informed choices consistent with their values.

Making tradeoffs in assessing quality

The provision of high technical care and high art of care can be examined from the perspective of an individual patient or of a population of patients. In many health systems in the world this distinction has become moot because, in reality, most doctors today are responsible for a group of patients.

We need to develop new clinical methods for use in day-to-day practice that incorporate a population perspective. An old study done in a poor area in the USA (the Appalachia region) many years ago illustrates this point. A general practitioner, who was the only doctor for more than 5000 people, would come to his office at 8.00 a.m. In the wintertime he might have more than 100 patients waiting to see him. He would ask all patients to sit down and then he would ask everyone who had a sore throat to stand up. Because it was the winter time, about one-third of the patients would stand up. He would then ask those patients who were allergic to penicillin to sit down. About 10% of the patients who were standing up would sit down. Following these two questions, he would ask all patients still standing to walk slowly to the right. His nurse was waiting, and right through their leather jackets or whatever outer clothing they were wearing, each patient would receive a shot of long-acting Penicillin. The doctor would then spend the rest of the morning seeing the patients that remained.

One of the early investigators in the field of quality assessment was somewhat horrified when he saw this process. But this doctor was practicing population-based medicine.

He was saying, 'I have two options. I can schedule appointment times and limit the number of patients who walk into my office. If I do that, many patients will get no care at all. However, I would have more time to provide better care to those patients who actually see me (if better care is correlated with more time with the physician, something that research studies have not been able to substantiate). On the other hand, if I develop a triage system based on scientific evidence, then perhaps I can practice good care for individuals and for a population at the same time'. The compromise he came up with follows many rules of good medicine. Perhaps its most serious drawback was putting holes in leather jackets.

This story illustrates that if we are serious about improving quality of care, we need to decide whether the group we are caring for is the population as a whole, or the patients that choose to see us at a point in time. This dilemma lies at the heart of the agony currently engulfing managed care in the USA as bureaucrats, administrators, and others attempt to increase physicians' list size, which in turn decreases the amount of time physicians have to spend with each patient. Non-physicians are trying to deliver population-based medicine within a defined budget. The leather jackets are objecting [6].

Medicine is full of procedures and tests that may produce slight marginal benefit or even be slightly harmful. No matter how much science we apply, we will never know for certain whether these services are slightly beneficial or slightly harmful. Thus a key policy question is whether society should pay for tests and procedures of marginal or no proven benefit when patients demand such interventions. Here are some telling examples.

Example 1. A man gets up two to three times in the middle of the night to urinate because he has an enlarged prostate. He wants to fix this problem and he is willing to take the slight risk of death, and the moderate risks of incontinence and impotence to do it. Should we pay for this procedure, which society would not judge to be cost-effective, out of either public money or publicly subsidized money?

Example 2. People should be able to choose a physician that they trust, but should they be able to choose the sex, race, age, perhaps even the hair and eye color of the doctor they see?

Example 3. What if we conclude, as research has already demonstrated, that sending reminder cards and having people call up in friendly voices will increase the number of people who get pap smears, mammographies, or blood pressure checks but that these activities increase total costs. How many reminder cards should be sent? How friendly is friendly enough? We do not have sufficient resources to do it all.

These vignettes lead us to make the following recommendations: (i) people's values vary regarding whether they want to maximize the technical or art aspects of care. In addition, it is very difficult to make rational choices under emotional stress. Thus we should make a great effort to reduce substantially the variation in quality of care (of course, without reducing the mean level of quality) so that people do not have to make these choices; (ii) we need to develop a clinical tool kit that will help physicians deliver high quality of care to populations that is also acceptable to individual patients. The clinical tool kit to do this has not been developed – its framework is barely visible, but we may need to fundamentally alter today's clinical process; (iii) we need to determine what people are willing to give up so that the proportion of the health system that is publicly funded or subsidized is compatible with the goals of society, fosters solidarity within the population, and does not destroy the pursuit of happiness in the name of health. If we want to make care more consistent, at a cost society might deem reasonable, we must forego receiving care of equivocal or marginal health benefit. We would also want the art of care to be high enough so that the patient, if she wishes, is empowered to be an equal partner in managing her own health.

We must find out what people value and figure out how to design an affordable health system that provides it. No health care system in the world is explicitly designed on this principle.

Choosing measures to assess quality

We can produce information about quality for public consumption that is based on structural measures (innate characteristics of physicians, nurses, the system), process measures (what health professionals do to people), or outcome measures (what happens to people, particularly in terms of their health). In an ideal world in which there was sufficient knowledge to predict with absolute certainty the relationship between structural, process, and outcome measures, one would pick measures of quality that could be obtained inexpensively and unobtrusively. Unfortunately, there are some real problems in using both structural and outcome measures versus process measures to assess quality.

Structural measures

Research has examined whether structural measures of quality predict what actually was done to patients. For instance, does a physician who is board certified produce better processes or outcomes than one who is not board certified? In general, relationships between structural and process variables are weak, inconsistent, and paradoxical [7]. Thus it is unwise to develop public information on quality of care that is based solely on structural measures. The following example demonstrates why.

The USA spends more money than probably any country in the world on hospital accreditation, yet study after study has demonstrated huge variations in quality of hospital care. There is no evidence to suggest that, in the absence of accreditation, the variation in hospital quality in the USA would be any greater [8]. The UK has a very sophisticated regionalized system of health care. Yet work we did in the UK in one region suggested that hospitals in the same region produced very different rates of the appropriateness of the use of coronary angiography. Measuring the amount of

regionalization is not, unfortunately, a very good measure of quality [9].

Outcomes

In some ways outcomes are also poor measures of quality of care. Health outcomes are exactly the kind of information people want when they select a provider or a hospital. However, outcomes are only partially produced by health services and are frequently influenced more by other factors, (e.g. natural history of the disease, patient physiologic reserve, or patient age). Consider, for example, a patient who comes into the emergency department with a heart attack. If one did nothing for this patient other than relieve his pain, he would be likely (60–70% probability) to leave the hospital alive and be able to resume his normal daily activities. Thus care that was absolutely atrocious from a medical standpoint would result in good outcomes for most patients. To use outcomes as a marker of quality, we need to adjust for differences in case mix and other external factors to ensure fair comparisons among institutions or physicians.

Using outcomes to measure quality is further complicated because many outcomes of interest occur years later, and thus are rendered useless as measures of quality for accountability. For instance, to compare quality of care for patients who have breast cancer, one might want to use an outcome such as the 5-year survival rate. By the time that information was available, it would reflect care that was actually given 7 or 8 years previously. During that period, the institutions that provided that care could have changed markedly. Thus an 8-year-old piece of data would say very little about the quality of care rendered to patients currently.

In addition, if we are serious about ever using outcomes to measure quality of care, we will need to address some uncomfortable issues: (i) we will have to require patients to provide data about what happens to them; and (ii) we will have to provide care in a regionalized manner so that outcomes can be accurately assessed. We will consider the first issue in the next section of this paper, where we discuss obtaining data to assess quality. We will begin discussion of the second issue by considering the operation carotid endarterectomy [10].

Carotid endarterectomy is an operation in which people, most of whom have had a transient ischemic attack, have an obstruction in that artery removed. When this process is done well, it decreases the likelihood of future strokes or death. But the operation is dangerous and can cause immediate death, heart attack, or stroke.

Carotid endarterectomy has been studied extensively in randomized controlled clinical trials that compared medical and surgical treatment [11,12]. However, the people who designed these trials chose surgeons and hospitals that they knew were good and had low complication rates and patients who had higher than average 5-year stroke risk and lower than average surgical risk. These studies showed that patients who had obstruction of the artery and a transient ischemic attack and who went to 'the surgeon selected to participate in the study' had better outcomes – lower rates of stroke and death – at 5 years than if they had had medical therapy. On the other hand, work we did demonstrated that the complication rate of this procedure, when performed by the average surgeon who does it, may be as much as threefold higher than that reported in the randomized controlled trial [13]. These higher complication rates suggest that medical therapy is better than surgical therapy.

Let us suppose that you have this condition and come to the first author as your internist to ask if he would recommend surgery or medical therapy. There is no current outcome reporting system anywhere in the world that would allow me to give you a database-derived answer to this question. Perhaps I should send you to one of the places that participated in this trial and hope they pay as much attention to you as they did when they were trying to prove that this surgery worked. Perhaps I can assume that because I am in an academic medical center, my surgeons are as good as the surgeons that participated in the trial. Or perhaps I should be more realistic and say, 'I do not know how to distinguish a good surgeon from an average surgeon, and therefore I am going to recommend that you have medical therapy'.

How could we alter this scenario? We would have to dramatically reduce the number of hospitals in the USA that are permitted to perform this operation. The operation works well if the total stroke and death rate is less than a few percent. If it is as much as 6 or 7%, then the operation, for most patients, is probably not worth doing. The difference between the operation that is worth doing and the one that is not worth doing is about three to five extra deaths per 100 people operated upon. To detect a statistically significant difference at this level in a valid way, we would need information on 1000 or more operations so that we could confidently say that an institution's death and stroke rate was low enough to make the operation worthwhile. There are about 100 000 of these operations performed each year in the USA. If we divide this 100 000 by 1000, then we would have only 100 hospitals in the entire country that would do this operation. If we believe that the number needed to obtain statistical validity is 2000 operations per hospital, then only 50 hospitals would be allowed to do this procedure.

If we do not go through a process like this one, we will never be able to help the patient who comes for advice. We will never know what is right to do. Thus before we design, conduct, and most importantly pay for randomized controlled clinical trials, we must be prepared to couple the results of randomized controlled clinical trials to policy recommendations about how care should be provided for a given procedure so that we can determine whether or not the successful arm of the randomized trial is worth providing in a community setting.

Process measures

Process measures are only as good as the evidence that associates them with improved outcomes. Process assessments produce the harshest judgment of the quality of care. In a comparative study of five different methods to assess quality, four of 300 patients received care that met all of the explicit criteria concerning process of care [14]. That

is, just four out of 300 patients got everything that they were supposed to get during the 6 months that their care was observed. If one assumes that high quality of care means that patients receive everything they should get, then 96% of people did not receive high quality of care. On the other hand, if one looked at outcomes that were potentially improveable, 25% of people had outcomes that could be improved by better process of care. Most, but certainly not all, people got well in spite of poor care. The medical profession makes many errors of both commission and omission and we often get away with it. Occasionally we lose.

For the vast majority of medical conditions, we will need to use process measures to assess quality. Thus the first priority of government in developing a quality tool kit should be to develop measures of the process of care. There are exceptions to this rule: when process and outcomes occur close together in time and when the process dominates the predictors of outcome. There are common operations, such as carotid endarterectomy and coronary artery bypass surgery, whose quality may be assessed by outcomes [15]. There are common hospitalizations that have high death rates, such as pneumonia, heart attack, heart failure, and stroke, where risk-adjusted models that compare hospitals by their death rates following hospitalization may be the best method to assess the quality of care patients receive. But, in general, process measures should be used to assess quality. We will need to work hard to make such measures understandable by all potential users.

In sum: regardless of what we would like to have happen, most of the quality indicators that we should use will be process based. When outcome indicators are needed, they will require changes in the way we deliver care so that numbers will be sufficiently large to make valid statistical statements. Whether this is acceptable in the USA is unknown. We should not fund clinical research studies if we do not have a strategy for coupling the results from those studies with a system to deliver services in a high quality manner.

We must provide a few additional comments about the use of outcome data. If the goal is to provide historical information about the progress, or lack thereof, that is occurring in a country's health care system, then outcome data may be very appropriate to use. Outcome data can also play an important role in evaluating changes in policy. For example, one could produce a time series that answers the following question: Is the survival rate from those cancers that are treatable, such as leukemia, lymphoma or breast cancer, increasing over time in the USA or in any country? Such a time series, even though the data in it may be 5–7 years old, would still provide interesting information about whether the US health care system was improving over time. Disaggregating such time series results by race, poverty status, and other important demographic characteristics might also provide insights into the equity of the health care system. Such information should be made available routinely in published reports about the quality of care in the USA or in any other country.

We used outcome data to evaluate the introduction of the Prospective Payment System in the USA, which overnight changed hospital reimbursement for Medicare patients from a fee-for-service, cost-plus basis to a fixed price for diagnosis related groups (DRGs) system [3]. The evaluation was based on abstracting data from a national sample of over 18 000 patient records and producing risk-adjusted mortality rates for patients who were admitted to hospitals before and after the institution of the DRG program with either pneumonia, stroke, heart attack, heart failure, hip fracture, or depression. We found, in general, that there was no change in the mortality trend line after DRGs were introduced. This was viewed as a positive finding: the program made it much easier for the federal government to control the amount of money spent on hospitals, but it did not, in general, affect the quality of inpatient care. The tragedy of this study is that there was no commitment on the part of the US government to repeat it to see whether the outcomes that occurred early in the implementation of DRGs were maintained into its maturity.

Sources of data for measuring quality

In part, the type of measure one uses to assess quality (i.e. structural, process, outcome) dictates the source from which data about quality should be obtained. There are multiple sources of such data, ranging from routinely collected data that are part of delivering health care, such as claims forms in a fee-for-service system, to data from patient surveys, medical records, or data obtained from direct observation of patients.

Each of these sources has its strengths and weaknesses. For instance, consider a patient who was being told she had breast cancer. The doctor may have told the patient: (i) you have breast cancer; and (ii) here are the various options that you have regarding treatment for this disease. This conversation could have been recorded on a tape or videodisk. However, the doctor may not have recorded any of the conversation in the medical record. Furthermore, the patient, when asked whether the physician informed her of treatment options, may have been so emotionally distraught that she did not remember the doctor's explaining her therapeutic choices. If a quality of care criterion was that the doctor should have a discussion with the patient about treatment choices, then the data obtained from both the medical record and from the patient would indicate that the doctor had failed. However, the data obtained from the audiotape would indicate that the doctor had complied with the criterion. What should be the standard? If the patient was in no frame of mind to actually hear what was being said, does the doctor get credit for saying it anyway? If the audiotape indicated that the doctor did inform the patient, should the doctor be given credit even though he did not record the information in the medical record?

Examples such as this one illustrate that different information is obtained from different sources. This does not mean that we will never know the truth unless we use multiple sources, multiple observers, and spend more money collecting data about quality than providing quality care. It does mean,

however, that we must assess the validity and reliability of data when assessing quality.

If the type of data to be used in assessing quality is carefully chosen, valid information about quality of care can be obtained. For instance, billing data can be used to assess whether or not a procedure is given to a patient. We used billing data almost one-quarter of a century ago to determine that many doctors used antibiotic injections inappropriately in children [16]. We also determined from billing data that not only did physicians use antibiotic injections inappropriately, but they also gave the wrong antibiotic – one likely to produce more harm than benefit. We used medical record data to determine how quality of care varied among patients hospitalized with heart attack, pneumonia, stroke, and heart failure [3]. Medical record data showed that when services that should have been given to patients – such as the right drug or the right procedure – were not given, then the probability that the patient experienced an abnormal event increased. In addition, data from patients can be used to describe how the patient felt about the care episode, what happened in terms of health and functional status, and whether or not the patient understood how to manage a chronic health problem [17].

Policy decisions can affect our ability to access the best source of data. Quality of care assessment can be extraordinarily expensive unless it is possible to use data sources, whether they are medical records or billing data, without asking for the patient's permission. There is currently a vigorous debate in the USA about the circumstances in which data contained in bills, in encounter forms, and in medical records can be used for purposes such as assessing quality of care [18–20]. If we make it difficult to obtain these records, and require patient permission, then assessing quality of care may become impossible to do reliably or validly.

In addition to obtaining information that is already maintained in some database or record, we must address under what circumstances one can obtain new data from a patient. For instance, suppose we want to know whether a patient has been satisfied with his or her most recent mental health service. Suppose we want to know whether a patient following a hip fracture and surgery can walk up a flight of stairs 6 months later. What human subjects review process do we have to go through to be allowed to collect those data and contact the patient?

There is no question that collecting data is an imposition on people. But to obtain some privileges, people in a developed society must provide information and give up some of their privacy. In the USA, one must report for jury duty when summoned. One must be fingerprinted, take an eyesight test, and pass a written examination to obtain a driver's license. As a geriatrician, the first author is required to report to the Department of Motor Vehicles (DMV) any patient who has Alzheimer's disease so that his license can be removed. An internist/neurologist must report to the DMV if a person has a seizure so that her license can be likewise restricted. How should we tradeoff the need for information versus the right to privacy in the quality of care area?

Our suggestion for compromise is as follows. Patients should be required to provide information about what happened to them, to complete questionnaires about what was done to them, and to allow people to abstract their medical records. Individually identified information must be carefully protected and violators prosecuted vigorously. However, the information must be used to improve quality and reporting it solely in peer-reviewed journals is not enough. Government, purchasers, providers, and consumer advocacy groups should be held accountable to improve care; otherwise why should patients give up their privacy?

We conclude this discussion about the tradeoff between quality assessment and confidentiality by considering how care could be assessed in a long-term care facility, i.e. a nursing home. Unless there are major scientific breakthroughs in the next few years, a large percentage, perhaps even a majority, of women in developed countries will spend some time in a facility that looks like, feels like, or smells like a nursing home. Perhaps as many as half of these women will not have even a single visitor per year or an advocate to help ensure that they are well treated. Many nursing homes are staffed by people who are paid minimum wage and rate working in these facilities slightly better than working in a prison.

How can we safeguard and improve the quality of medical care given to patients under these circumstances? Let us consider a patient who was too weak to turn herself in bed and should be turned gently by a nurse every 2 hours; or a person who is too weak to feed herself and should be fed in a humane way at least three times a day. If we developed quality of care criteria to see that these activities happened, and if we used nursing home medical records to assess compliance with those criteria, it is likely that in many nursing homes we would stimulate an atmosphere of increased fraud in which more time would be spent documenting things that were not done than doing them.

Another possibility would be to put a video camera on every patient's bed with a system that allowed somebody to look at a random set of the videotapes to determine whether the behaviors we wished to encourage actually took place. Do we need patient permission to implement such a system? Is the system a violation of modesty – and of many other things as well? Suppose we gave women in a nursing home a choice about having a camera on their bed and using data from that camera to assess their quality of care. If only some women chose it, would we find that the camera was such a powerful mechanism in influencing quality that those who did not have a camera would never get any attention at all? If the majority of people agreed that in the name of safety, cameras were necessary on all nursing home beds, would we be willing to implement such a policy?

We believe we must confront such issues. We also believe that in most cases we will decide that the price we are paying in terms of relinquished confidentiality is more than worth the improvement in quality of care that will result when information about quality becomes routinely available in all settings.

Equity in reporting quality of care

The way we measure quality must be fair to providers. Suppose we observe a difference in either processes or outcomes of care between two providers (whether they are hospitals, physicians, or nursing homes), and we wish to do something with this information. 'Doing something' may involve anything ranging from making the information available publicly to using it to regulate or license, or even in some cases giving feedback to the provider who produced the information.

In most cases, providers react defensively to information that suggests deficits in quality. 'I cannot be that bad. The data must be wrong.' In some instances, the pressure from providers may become so great that the mechanism for assessing quality of care is defeated. This is what occurred in the USA with the public release of hospital mortality data by the Health Care Financing Administration (HCFA). The data that were released were adjusted for severity at the time of hospital admission, but the adjustment included only clinical information collected routinely from hospitals at the time patients were discharged. This information contained the procedures that were done, the diagnoses the patient had, and certain demographic variables. These variables were all used in logistic regression models to adjust for case mix; after adjustment, observed and expected death rates were published for all US hospitals.

Certain hospitals, especially some in the inner city, were found to have much higher adjusted death rates than expected. The administrators of these hospitals complained to the HCFA administrator that the ratings were unfair because their hospitals had sicker patients and the method of case adjustment was not sufficiently sensitive to capture that fact. In fact, this may have been true. Work we did in New York City comparing outcomes of care in public versus private hospitals showed that the method of adjusting the data strongly influenced decisions about which hospitals provided better or worse care [21].

The HCFA administrator had two choices. He could pay to collect more detailed clinical data and thus produce better but still not perfect mortality models. If he did this, he could limit the patients whose mortality rates would be publicly released to those with one of a few diseases (pneumonia, stroke, heart attack, and heart failure) that make up a large majority of the hospital deaths for people over the age of 65 years. While developing this technology, he could have continued to report existing mortality data. However, he decided to scrap the whole system. Why? That is a difficult question to answer. Collection of the additional clinical data would have cost an extra $30–50 per patient discharged with those conditions from the hospital. Perhaps that money would have been seen as wasted and better spent on something else – perhaps providing more medical care or more tanks.

This dilemma arises repeatedly in the quality of care field. People want valid and reliable measures of quality of care but they do not want to pay for them. They vigorously oppose any system that has demonstrable error in it that could be improved with a better data collection system because it is not fair to risk the reputation of a single hospital or a single doctor through bad data about quality. On the other hand, when one proposes a more detailed clinical data collection system, the complaints are that it costs too much, is not feasible, and probably is not that important.

We have to develop a set of policies that will resolve this impasse. There is some evidence that might help in this regard. For example, in a study of prenatal care that used process measures, we found that it did not matter whether one adjusted for case mix in comparing performance among six health management organizations [22]. The comparisons were just as valid with or without adjustment for differences in case mix. This is generally the case in making process comparisons. In addition, for procedures such as coronary artery bypass surgery, where patients are usually operated on after being stabilized, adjustment for differences in case mix when using death rates as the quality measure, even at a detailed clinical level, does not contribute as much as it would if one adjusted for differences in case mix among patients admitted with a medical condition where the severity of the admission for heart attack, pneumonia, or stroke may differ greatly as a function of the hospital to which the patient is admitted [23].

If we are serious about reporting information about quality, we will need to determine the sensitivity of our results (e.g. labeling a good hospital as a bad hospital) to measures based on inexpensively collected data as opposed to more expensively collected data. Using that information, we need to make a decision about which data system should be used to assess quality of care in a valid and reliable way so that it is judged to be fair enough (not perfectly fair because there is no perfect system) to the doctors and hospitals included in the assessment. It is now possible to do this for many chronic conditions and for many procedures.

Looking to the next century

It would be inappropriate to end this paper on a pessimistic note. A great deal of research has been devoted to measuring the appropriateness and necessity of health care. This work has used a method that combines a review of the scientific literature about the effectiveness and efficacy of a procedure with evidence obtained from a multi-specialty group judgment process [24]. This work has shown that in many countries in the world, a large percentage of procedures are not needed, while at the same time procedures that are needed to improve health are not offered to patients. Research has also shown that physicians fail to ask patients many clinical questions that are vital to deciding whether to do a procedure; or if patients are asked, it occurs in such a hurried and informal manner that valid and reliable information is not obtained. This is a definition of chaos.

The appropriateness technology is sufficiently well developed that it could replace the current way of practicing medicine. There is no question that for the 70–100 procedures that make up most of what we do, we should have a computer-based, prospective system by which we make sure

that when we, as physicians, treat patients, we ask the questions required to decide whether to do the procedure. We should make sure that the patient verifies the responses to those questions. We should combine the answers from patients with the results of tests into an appropriateness and necessity score that provides an initial assessment of the appropriateness and necessity of the procedure.

This real time assessment would not be used as an absolute rule to determine whether or not the procedure would be done. But it would serve as a starting point for negotiation between patients and doctors as they go about the business of deciding what should be done. If the assessment is overruled by either the doctor or the patient, then a simple clinical justification for that decision would be provided and used to improve, if deemed correct, the clinical system.

The appropriateness technology is feasible to use today and could radically improve the quality of care provided, at least for expensive procedures. It would help ensure that people get necessary things when they need them and do not get things they do not need.

In addition, advanced tools are also currently being developed that would allow a more valid, reliable, and comprehensive assessment of the quality of primary care across multiple conditions [25–27]. These tools are based upon data obtained both from the patient's medical records and from the patient. Information could be reported at the medical group level about the quality of primary care that is currently being given. Because these tools contain many indicators across multiple conditions and collect data from both records and patients, they are more likely to be valid and reliable and less likely to be gameable (i.e. less likely to be subject to inappropriate manipulation on the part of the provider). For example, because the quality for over 70 conditions is assessed, it would be virtually impossible for medical groups to reallocate funds from unmeasured to measured conditions so that their quality score would increase. These tools could be used to produce a comprehensive assessment of the quality of primary care at the country, region, and medical group level.

The last third of the century has produced information that could be used today to measure and improve quality of health care that is being provided in the western world. It would be a shame not to take advantage of this information, much of which was publicly funded. Right now we are comfortable spending $100 000 or more to extend the life of one patient by one year. We can get more value (i.e. health) for the money we spend if we invest more at the margin in a consistent and stronger effort to measure quality of care and improve it than if we provide more care to insured middle-class Americans.

References

1. Brook RH, McGlynn EA, Cleary PD. Measuring Quality of Care. *N Engl J Med* 1996; **335:** 966–970.
2. Schuster MA, McGlynn EA, Brook RH. How good is the quality of health care in the United States? *Milbank Q* 1998; **76:** 517–563.
3. Kahn KL, Keeler EB, Sherwood MJ *et al.* Comparing outcomes of care pre- and post-implementation of the DRG-based prospective payment system. *J Am Med Assoc* 1990; **264:** 1984–1988.
4. Oxman AD, Thomson MA, Davis DA, Haynes RB. No magic bullets: a systematic review of 102 trials of interventions to improve professional practice. *Can Med Assoc J* 1995; **153:** 1423–1431.
5. Lohr K, Brook RH, Kamberg C. Use of medical care in the RAND Health Insurance Experiment: diagnosis- and service-specific analyses in a randomized controlled trial. *Med Care* 1986; **24:** S1–S87.
6. Brook RH. Managed care is not the problem, quality is. *J Am Med Assoc* 1997; **278:** 1612–1613.
7. Brook RH, Park RE, Chassin MR. Predicting the appropriate use of carotid endarterectomy, upper gastrointestinal endoscopy, and coronary angiography. *N Engl J Med* 1990; **323:** 1173–1177.
8. Keeler EB, Rubenstein LV, Kahn KL *et al.* Hospital characteristics and quality of care. *J Am Med Assoc* 1992; **268:** 1709–1714.
9. Gray D, Hampton JR, Bernstein SJ, Brook RH. Clinical practice: Audit of coronary angiography and bypass surgery. *Lancet* 1990; **335:** 1317–1320.
10. Brook RH. Adapting practice patterns to a managed care environment: carotid endarterectomy – a case example. *J Vasc Surg* 1996; **23:** 913–917.
11. Hobson RW, Weiss DG, Fields WS *et al.* Efficacy of carotid endarterectomy for asymptomatic carotid stenosis. *N Engl J Med* 1993; **328:** 222–227.
12. North American symptomatic carotid endarterectomy trial collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *N Engl J Med*, 1991; **325:** 445–453.
13. Winslow CM, Solomon DH, Chassin MR *et al.* The Appropriateness of performing carotid endarterectomy. *N Engl J Med* 1988; **318:** 721–727.
14. Brook RH, Appel FA. Quality of care assessment: Choosing a method for peer review. *N Engl J Med* 1973; **288:** 1323–1329.
15. Brook RH. Health care reform is on the way: Do we want to compete on quality? *Ann Intern Med* 1994; **120:** 84–86.
16. Brook RH, Williams KN. Effect of medical care review on the use of injections. *Ann Intern Med* 1976; **85:** 509–515.
17. Hays RD, Brown JA, Spritzer KL *et al.* Member ratings of health care provided by 48 physician groups. *Arch Intern Med* 1998; **158:** 785–790.
18. Detmer DE. Your privacy or your health – will medical privacy legislation stop quality health care? *Int J Qual Health Care* 2000; **12:** 1–3.
19. Van den Hoven J. Privacy and health information: the need for a fine-grained account. *Int J Qual Health Care* 2000; **12:** 5–6.
20. Willison D. Privacy and confidentiality concerns – are we up to the challenge? *Int J Qual Health Care* 2000; **12:** 7–9.

21. Shapiro MF, Park RE, Keesey J, Brook RH. The effect of alternative case-mix adjustments on mortality differences between municipal and voluntary hospitals in New York City. *Health Serv Res* 1994; **29:** 95–112.

22. Murata PJ, McGlynn EA, Siu AL *et al.* for the HMO Quality Care Consortium. Quality measures for prenatal care: a comparison of care in six health care plans. *Arch Fam Med* 1994; **3:** 41–49.

23. Williams SV, Nash DB, Goldfarb N. Differences in mortality from coronary artery bypass graft surgery at five teaching hospitals. *J Am Med Assoc* 1991; **266:** 810–815.

24. Brook RH, Chassin MR, Fink A. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care* 1986; **2:** 53–63.

25. Schuster MA, Asch SM, McGlynn EA *et al.* Development of a quality of care measurement system for children and adolescents: methodological considerations and comparisons with a system for adult women. *Arch Pediatr Adoles Med* 1997; **151:** 1085–1092.

26. Malin JL, Asch SM, Kerr EA, McGlynn EA. Evaluating the quality of cancer care: development of cancer quality indicators for a global quality assessment tool. *Cancer* 2000; **88:** 701–707.

27. McGlynn EA, Kerr EA, Asch SM. A new approach to assessing the clinical quality of care for women: the QA tool system. *Women's Health Issues* 1999; **9:** 184–192.

Appendix: Quality factoids from work at RAND

1. RAND in the 1970s performed a $100 million community-based randomized controlled experiment to test (in the fee-for service system in the USA) the effect of cost-sharing, (i.e. deductibles and co-insurance versus free care), on use, cost, and health [1,2]. The conclusions from this 5-year landmark study, in general, were that for the average adult and child, cost-sharing, although reducing use by one-third or more, resulted in no change in health status. This lack of an impact on health was at least partly due to the fact that neither compliance with quality of care criteria, nor the likelihood that a patient-initiated episode of care would be one for which medical care was judged to be highly effective as opposed to not effective at all was changed by having free care. Thus, a financial incentive such as free care did not specifically increase the use of health care for medically necessary reasons. Health care use with free care was 30% higher for both effective and ineffective patient care episodes. Thus, a patient in a cost-sharing plan was equally as likely not to show up in the doctor's office for a symptom complex such as a cold for which medical care is not effective as she was likely not to show up for an ulcer, for which therapy could be effective. Furthermore, when the patient did show up, she received about 60% of the services she needed regardless whether care was free or not. This occurred at a time when a US physician could do anything he desired and there was no concept such as utilization review.

2. The conceptual framework for assessing an array of disease-specific outcomes was developed over 20 years ago [3]. This framework was illustrated for a child with asthma. Asthma is the most important common chronic condition for children. To perform a comprehensive assessment of the outcome of asthma care and to use that assessment to determine if a physician provided optimal asthma care would require the following steps:

(a) First, all patients with asthma would need to be identified and their illness severity determined. Children who have more severe asthma may be triaged either to medical groups or physicians who are known to be experts in treating such children. However measuring severity is not straightforward. Based on a review of the literature and a consensus panel, it was determined that very little research was available to decide how to measure illness severity at the time a child with asthma first presents to a medical group. One of the conclusions of the research study was that more research was needed in this area but research in this area remains a low funding priority.

(b) Second, the outcomes believed to be important are identified and include the following:

(i) disease severity
(ii) amelioration of symptoms
(iii) patient understanding of his or her disease
(iv) disability (work/school lost days)
(v) loss of sleep
(vi) school/work performance
(vii) patient anxiety
(viii) patient depression
(xi) patient feeling of inferiority
(x) patient fear of shortened survival
(xi) parent/patient expectations for the future
(xii) family anxiety
(xiii) family depression.

Each of these outcomes was considered by our expert physicians to be important. A review of the literature as well as the opinions of an expert clinical panel indicated that there was little research that would help guide the development of models to explain these outcomes as a function of either innate characteristics of the patient and his or her family, or whether the quality of medical care was good or bad. What was clear is that a different multi-variable model was required to predict the occurrence of each outcome. Further, doctors who are good at preventing the patient from losing sleep because of asthma symptoms may not also be good at helping the child understand or reduce her fear of shortened survival or recognizing or treating the family anxiety and depression that results from having a child with asthma.

This work concluded that there has not been, for even a single chronic condition, a comprehensive examination of the link between what we do in medicine and the areas of important outcomes, let alone the use

of that information to say something about quality of care. Currently there is either no adjustment for differences in severity of illness when outcomes across medical groups or doctors are compared, or only one disease-specific outcome is used, such as mortality, or a general health status measure is used as the outcome measure and this measure does not capture adequately all the dimensions listed above. More research is needed to determine how to use the outcome assessment method to make a definitive statement about the quality of care given to a patient with a specific disease.

3. Five quality of care assessment methods were compared to evaluate the care received by patients with one of three conditions: ulcer, hypertension, or urinary tract infection [4]. The five methods were: implicit process judgment, implicit outcome judgment, implicit quality of care judgment, explicit process judgment, and estimation of group outcomes. These are the fundamental methods that are still used today to assess quality of care.
   (a) The implicit process method requires a health professional to form a judgment about whether what was done to a patient is adequate or inadequate, based on a review of the medical record.
   (b) The implicit outcome method requires the health professional, without the use of any explicit criteria, to read what was done to the patient and what happened to the patient. To obtain the outcome information, a standardized outcome assessment some time after care was received is performed. The health professional is then asked the question, 'Considering what was done to the patient and what happened to the patient, could the outcome or health status of the patient been improved?'
   (c) The implicit quality of care judgment provides the health professional with both what happened to the patient and what was done to the patient, and asks the physician or health professional to rate the quality of care, considering both the process and outcome of care.
   (d) The explicit process judgment requires comparing what was done to the patient with explicit, previously developed criteria that are based on a review of the scientific literature and expert judgment.
   (e) The fifth method compares what happens to a group of patients whose demographic and clinical characteristics have been precisely defined against explicit external benchmarks or a group judgment of what should be the outcome of care.

   When these five methods that form the basis of all modern methods of quality of care assessment were compared, it was shown that about 2% of the patients met all of the explicit process criteria, that the process of care when judged implicitly was adequate for about one-quarter of the patients, and that the implicit outcome judgment was unimproveable for two-thirds of the patients. The results of the implicit quality of care review were very similar to those from the implicit process judgment.

   This study has two very important implications for policy. First, any process judgment of quality of care will produce a much harsher assessment of quality than will an outcome judgment. This is true because people can get bad care and still survive and prosper. Second, physicians believe that a quality of care assessment should be based more on what was done to the patient than what happened to the patient and that outcome data are almost irrelevant to a physician's judgment of the quality of care rendered to a given patient.

4. A quarter of a century ago we demonstrated that routinely collected claims data from a fee-for-service system could be used to make powerful statements about quality of care and that these claims data, in turn, could be used to monitor the results of a program directed at changing the use of injections in the ambulatory arena [5,6]. In the Medicaid Program in New Mexico we observed that about two out of every five ambulatory visits ended with an injection, nearly 50% of the injections were antibiotics, and that most of the injections were for the wrong antibiotic. This determination was made possible by linking claims data that contained the patient diagnosis with procedure data that contained information about whether an injection was given and if so, what medication was used. Based on the above findings, a dual strategy was implemented to both educate physicians about what is the appropriate use of antibiotic injections and to deny payment for injections that were felt to be medically inappropriate. The injection rate fell quickly from about two out of five to one out of seven visits ending in an injection, and most of the injections that were given were ones that were medically appropriate. This study indicates that valid, routinely collected data can be used to both measure and improve quality of care and that before more complex methods of quality of care measurement are used, secondary data bases should be exploited for the impact that they potentially could have on understanding the current level of quality of care and how that care might be improved.

5. In another study done on psychosocial problems and chronically ill children over 25 years ago, 44 children with a chronic illness and their parents were interviewed concerning their expectations for the current visit [7]. The doctor/patient interaction was tape-recorded. Identical categories of information were extracted from the tape recording and from a review of the patient's medical records. Examples of findings from this study are: (i) although parents expected that about three-quarters of the psychosocial aspects of care should be covered by the doctor, only on-quarter were actually discussed in the visit; (ii) unfulfilled expectations were associated with a lower satisfaction with the medical care received; and (iii) doctors recorded about 80% of the discussions of symptoms and physical findings in the patient's medical record, but only 25% of the

discussions of psychosocial problems. This small study, which has been confirmed by many others, indicates that different aspects of the patient's problems are more or less likely to be covered in a routine visit. Patient expectations are not met very often in the psychosocial dimension of health care. Quality of care assessment based on a written medical record is more valid when one examines information about whether history and physical items and symptoms were assessed as opposed to whether psychosocial aspects of the patient's problem were discussed. This study suggests that patient surveys be used when one wants to assess the psychosocial aspects of health care and that the clinical process be re-engineered so that the doctor is aware of the patient's expectations for the visit before the visit actually begins.

6. It is tempting to examine the rate of use of a procedure across geographic areas and conclude that differences in the use of a procedure are related to the appropriateness of the use of a procedure. This was not found to be the case in a study of three procedures – angiography, carotid endarterectomy, and upper gastrointestinal endoscopy – in the USA [8]. These procedures were performed two to three times as frequently in the high-use area compared to the low-use area. However, differences in appropriateness of care by geographic area were small and could not explain the huge differences in the use of the procedure. This work and other work has confirmed that under- and overuse of services exist simultaneously in geographic areas and that reliance on secondary data sources about level of the use of a procedure, unless it is related in an explicit way to appropriateness or need, is a poor and potentially misleading marker of appropriateness or quality of care. This finding has been confirmed in small area analyses as well as in large area analyses and makes it virtually impossible to conclude anything about quality from variations in utilization rates [9].

7. One of the most important questions regarding using secondary data to measure differences in hospital outcomes is the validity of the method that is used to adjust for differences in case mix. We demonstrated, in New York City in the early 1980s, that conclusions regarding the quality of hospital care provided in municipal hospitals versus that in voluntary hospitals are dependent on the adjustment method used 10,11]. We developed logistic regression models for pneumonia, stroke, head trauma, hip repair, and myocardial infarction and compared the additional deaths potentially caused by poor quality of care in city hospitals versus voluntary hospitals under several models. We found that estimates of mortality differences between New York City municipal hospitals and voluntary hospitals are substantially affected by which secondary diagnoses are used in the case mix adjustment method. For instance, if we put into the model only patient characteristics such as age, sex, and those diagnoses that are almost certainly present on admission (such as diabetes) and not potentially caused by poor quality of care during the hospital stay (such as cardiac arrest), then there were 3.3 additional deaths per 100 people admitted in city hospitals for heart attack, 1.2 for pneumonia, 8.3 for stroke, 2.8 for head trauma, and 0.8 for hip repair. However, when we added all the diagnoses into the model that appear on the face sheet of the patient's hospital medical record and therefore are contained in the administrative data base maintained by the State of New York, only differences for stroke remained significant between the two hospital groups. The full adjustment model contains diagnoses that could either have been caused by bad quality of care, such as a patient who developed heart failure because his heart attack was inappropriately managed, or could have been present at time of admission and be a marker of patient severity with which the hospital had to cope. Recently, hospital diagnoses are being coded with an additional digit (in the States of California and New York) that indicates whether the co-morbid diagnoses were present at time of hospital admission or occurred during the hospitalization. Hopefully the use of this digit will increase the validity of case-adjustment methods using secondary data and may make it possible to compare mortality across hospitals for selected conditions without collecting additional clinical data. One of the most important research studies to be immediately performed should be to compare assessments of quality based on those systems that contain this added digit in the secondary database with systems that re-abstract the medical records to collect more detailed clinical data. If it is possible to develop valid risk-adjusted measures of hospital mortality for common conditions from secondary data that include this extra digit, then this could go a long way to making valid information public about differences in hospital quality.

8. Explaining differences in hospital quality has been difficult [12]. What has been established is that hospitals that have better processes of care have better outcomes. In a national study of almost 400 hospitals in the USA in the 1980s, people who received processes of care ranked in the top 25% for conditions such as heart attack, pneumonia, stroke, and heart failure had three to five fewer deaths per 100 people admitted to a hospital than those people who received care in the lower 25th percentile of the quality distribution [13]. This is an enormous difference, and because these data come from a nationally (USA) representative sample of hospitals, the results can be generalized and can be used to assert that hundreds of thousands of excess deaths are being produced in the USA because of variation in the quality of hospital care. There has not been a lot of work that has examined how these deaths could be prevented. However, in another study it was shown that the reasons for possible preventable deaths cut across many different clinical processes such as inadequate treatment of angina, inadequate fluid management, failure to control arrhythmias, inadequate

management of sepsis, inadequate diagnostic work-up, or inadequate airway or oxygen management [14,15]. What this means is that in order to fix this problem of preventable deaths, it will not be possible just to develop an automated drug dispensing system, but rather what one will need is a systems approach to improving and re-engineering hospital care where all aspects of the care process are improved. Targeting one aspect of the care process and hoping to make a major dent in the preventable deaths that occur in American hospitals will just chip away at the problem. Although the above findings regarding variation in hospital quality are accepted, it is also clear that being extremely confident about whether a specific hospital is a statistical quality outlier is difficult to establish. Although it is unlikely that a hospital that provides care in the bottom third of the quality distribution will, if a different method is used to measure quality, move to the top third of the distribution, it is clear that the adjustment method used to correct for differences in case mix when assessing differences in quality of hospital care can change whether a hospital's ranking is significantly different from some pre-established cutpoint.

Thus, what has been learned in the field of measuring hospital quality using outcomes is that a few diseases produce most of the hospital deaths. There is large variation among hospitals in the quality of care for people hospitalized for these conditions. This variation results in a large number of lives (certainly over 100 000) being shortened in the USA today. There are many reasons for these preventable deaths rather than one or two reasons that could be easily identified and corrected with simple management systems. Predicting precisely whether a given hospital is a statistical outlier is difficult because of variation in performance over time, relatively small sample sizes, and the methods used to adjust for case severity. Nonetheless, hospitals that are at the bottom of the quality distribution are unlikely to move to the top of the distribution if the adjustment method is changed and vice-versa.

9. A great deal of work has been done on measuring the appropriateness of medical care. If the health benefit exceeds the health risk by a sufficiently wide margin, then the procedure is defined as appropriate and is worth doing. This definition has been extended to measuring the necessity of care, which is a subset of appropriateness and means that if physicians did not offer patients an appropriate procedure, they would feel so upset they would want to go on strike or consider it morally reprehensible that they could not provide patients this service. Appropriateness studies have been done in many countries in the developed world using a method that is based upon reviewing the scientific literature and filling in the holes in the scientific literature with expert judgment. From this work there are many important generalizations. First, appropriateness of care varies remarkably from hospital to hospital, even in settings in which there is a national health system and regionalization. In one region in the UK, the appropriateness of the use of coronary angiography varied from 37% in one hospital to 63% in another hospital. In general, when appropriateness has been rated by panels in different countries in the western world, there is remarkably good agreement among them, and doctors in the western world seem to be using similar paradigms in judging appropriateness of the use of procedures [16,17]. The one minor exception to this might be physicians in at least one region, i.e. Trent in the UK. Results from both a US panel that rated appropriateness of care for coronary angiography and coronary bypass surgery and from a Trent regional panel were independently applied to actual patients who had undergone coronary angiography in Trent. For coronary angiography, if the US panels ratings were used, then 71% of the angiographies in Trent were judged appropriate; but if the Trent panel's ratings were used, then 49% of the same angiographies were judged appropriate. The difference in these figures was due to the fact that the Trent physician panel members were much more likely to base their ratings on the scientific evidence reported in randomized controlled clinical trials. When there was evidence from randomized controlled clinical trials regarding the efficacy of a coronary angiography, both the US and UK panels agreed that the procedure was appropriate to do. However, when that type of rigorous evidence was missing and was replaced with softer evidence from either descriptive studies, quasi-experimental studies, or even in some cases, animal models, the US panel might indicate it was appropriate to do the procedure, but the UK panel would rate such an indication as equivocal or inappropriate. Of all the physician groups we have examined, which include the Canadians, the Swiss, the Swedish, the Dutch, and the Israelis, the Trent UK panel was more true to the belief that a procedure should be labeled appropriate only if good evidence as opposed to judgment was available to support its use. With this caveat in mind, we have found consistent results regarding appropriateness across many countries.

For instance in four hospitals in Israel, the appropriateness of cholecystectomy varied from 64% to 83% by hospital and was not correlated with the geographical use rate of the procedure in the area served by the hospital [18]. In the USA we found that we could explain 4% or less of the variability in the appropriateness of care on the basis of standard, easily obtainable data about the patient, the physician or the hospital [19]. The characteristics examined were age, sex, race of the patient, physician age, board certification status, and experience with the procedure, and whether the hospital was a teaching one, a profit-making one, and its size. There were a few factors that were significant in predicting the appropriateness of the three procedures (upper gastrointestinal endoscopy, coronary angiography, and carotid endarterectomy). Performance in

a teaching hospital increased the likelihood that the procedure would be clinically appropriate. Angiographies were more often performed for appropriate reasons in older or more affluent patients. Being treated by a surgeon who performed a higher rather than a lower number of procedures, however, decreased the likelihood of an appropriate carotid endarterectomy by one-third from 40% to 28%. The latter finding is an ironic one, in that in order to do the procedure well and justify its use, you need to perform a lot of them. But obviously, one way of performing a lot of them in a market characterized by perhaps too many surgeons who can do this procedure is to perform more of them for reasons that are not appropriate in the first place. In a study of the appropriateness of hysterectomy in managed care plans (seven managed care plans in the USA) we found variation in the rate of inappropriate use of hysterectomy in these plans; 16% of women underwent hysterectomy for reasons judged to be clinically inappropriate, and almost two out of five women had the procedure for reasons that were judged to be either equivocal or medically inappropriate [20]. The appropriateness method has also been applied to identify underuse of care. In six hospitals in Los Angeles, an area in which there is a very high rate of the use of coronary angiography and coronary revascularization, we found that underuse of needed coronary revascularization occurred in 25% of people [21–23]. In other words, in the population of patients who underwent coronary angiography, based on findings from the coronary angiography, and the patient's history and symptoms, only 75% of patients actually had a necessary coronary revascularization procedure. Those people that did not get a coronary revascularization procedure died at much higher rates. Similarly, we found that only 43% of patients in the same geographic area who met necessity criteria based on the results of their history and cardiovascular stress tests actually received a coronary angiography within 3 months of the diagnostic test; 56% received the procedure within 12 months of the stress test. In Switzerland, in a study of 20 primary care physicians and over 7000 patient visits, we found that of the 611 patients who complained of upper digestive tract symptoms, underuse of endoscopy was identified in about 12% of these patients. Thus, findings from this method in multiple countries have demonstrated [24]:

(a) that physician panels around the world can be convened to develop appropriateness criteria;

(b) that these appropriateness criteria can be used to evaluate patient care;

(c) that when they are applied, both overuse and underuse can be detected;

(d) that underuse and overuse of the same procedure occur simultaneously in the same geographic areas, the same hospital, and in care provided by the same physician;

(e) that underuse and overuse occur regardless of the financial incentives that are operating;

(f) the underuse and overuse cannot be predicted from readily available structural data.

Thus, if we are going to improve the appropriateness with which services are used, we must tackle this problem directly, measure it prospectively, and use the data to improve the care process.

10. Results of assessment of the quality of prenatal care in six managed care plans in the USA produced the following conclusions [25]:

(a) that quality of prenatal care varies across the plans. This is similar to findings that have been mentioned previously;

(b) that on average, about 80% of routine screening tests and other routine prenatal care processes were done for women who were pregnant and treated in these health maintenance organizations but only 70% of the interventions to follow-up on abnormal findings or manage complications of pregnancy were done. One plan provided 95% of the routine screening tests while another plan provided less than 66% of them. The assessment of quality was based on information obtained from the medical record and demonstrated that, even in an environment in which population-based medicine is supposed to be practiced and care is managed, a large percentage of the services and tests that need to be done are not done.

11. A recent study demonstrated that it is possible to assess the satisfaction of patients with their physician group [26]. In a study of over 7000 patients and 48 medical groups, large variations in scores on multiple satisfaction scales were found. For instance, on the quality of care scale, which had a mean of 45 and a standard deviation of 10, the low plan had a score of 28 and the high plan had a score of 68. Negative patient ratings of care were significantly related to intention to switch to another physician group, difficulty in getting appointments, lengthy waiting times in the reception area and in the examination areas, the inability to receive consistent care from one physician for routine visits, and not being informed by the office staff when there was a delay in seeing the primary care physician. This study goes a long way to providing evidence that monitoring health care quality at the physician group level is possible and could be used for benchmarking, internal quality improvement, and for providing information to the public about how physician groups will meet their needs.

12. How do we put all these assorted facts about quality together in an attempt to improve the quality of care for a procedure such as carotid endarterectomy? [27, 28] What is needed is for this procedure to be used appropriately and performed in a technically outstanding manner.

(a) Up-to-date, explicit, clinically detailed, multi-specialty criteria need to be developed and used prospectively to decide under what circumstances carotid endarterectomy is both appropriate and necessary.

(b) A system to publicly report outcome data by physician and hospital must be developed because how well this procedure is performed must be the primary criterion in determining whether it should be used.
(c) To produce statistically valid outcome data (stroke and mortality results following carotid endarterectomy) we need to reduce the number of places where the procedure is done. In the USA about 100 000 carotid endarterectomies are performed per year. Obtaining statistically valid outcomes means that no fewer than 500 or 1000 carotid endarterectomies should be performed at any institution that does this procedure. At most, 200 hospitals in the USA should be performing carotid endarterectomy. The country would probably be even better off if only 100 facilities performed this operation.
(d) The appropriateness of the procedure should be explicitly measured and compared with the clinically detailed guidelines. It is vital that for risky procedures such as carotid endarterectomy, we know that they are being performed only when medically justifiable clinical indications are present. Before the procedure is performed, the physician should record symptoms, signs and needed treatments, and the results of tests and should check these against the appropriateness criteria to make sure the patient has a clinical condition that benefits from the operation; if this comparison needs to be overruled the clinician should state in the medical record the reason for this decision and it should be explained as well to the patient.
(e) Because of the variability in reading the results of the carotid angiography, which serves as the basis for doing a carotid endarterectomy, the reliability of angiography readings must be improved so that they can correctly identify those patients who will benefit from carotid endarterectomy. This could be done through slipping standardized previously assessed angiographies into the daily work to make sure that whoever reads the results of the angiogram is capable of rendering an opinion as to whether sufficient disease is present to warrant an operation. In addition, it may be impossible to assure adequate reliability with only a single reading; two or three independent readings may be needed to increase reliability to an acceptable level.
(f) Because of financial pressures that exist in all parts of the world to reduce the use of medical care, a system should be developed to make sure that people who actually need a procedure and meet necessary criteria for it actually are offered it.

References (Appendix)

1. Brook RH, Ware JE, Rogers WH. Does free care improve adults' health?: Results from a randomized controlled trial. *N Engl J Med* 1983; **309:** 1426–1434.
2. Lohr K, Brook RH, Kamberg C. Use of medical care in the RAND Health Insurance Experiment: diagnosis- and service-specific analyses in a randomized controlled trial. *Med Care* 1986; **24:** S1–S87.
3. Brook RH, Avery AD, Greenfield S. Assessing the quality of medical care using outcome measures: An overview of the method. *Med Care* 1977; **15** (suppl).
4. Brook RH, Appel FA. Quality of care assessment: Choosing a method for peer review. *N Engl J Med* 1973; **288:** 1323–1329.
5. Lohr KN, Brook RH. Quality of care in episodes of respiratory illness among Medicaid patients in New Mexico. *Ann Intern Med* 1980; **92:** 99–106.
6. Brook RH, Williams KN. Effect of medical care review on the use of injections. *Ann Intern Med* 1976; **85:** 509–515.
7. Lau RR, Williams HS, Williams LC *et al.* Psychosocial problems in chronically ill children: physician concern, parent satisfaction, and the validity of medical records. *J Commun Health* 1982; **7:** 250–261.
8. Chassin MR, Kosecoff J, Park RE *et al.* Does inappropriate use explain geographic variations in the use of health care services? *J Am Med Assoc* 1987; **258:** 2533–2537.
9. Leape LL, Park RE, Solomon DH *et al.* Does inappropriate use explain small-area variations in the use of health care services? *J Am Med Assoc* 1990; **263:** 669–672.
10. Shapiro MF, Park RE, Keesey J, Brook RH. Mortality differences between New York City municipal and voluntary hospitals, for selected conditions. *Am J Public Health* 1993; **83:** 1024–1026.
11. Shapiro MF, Park RE, Keesey J, Brook RH. The effect of alternative case-mix adjustments on mortality differences between municipal and voluntary hospitals in New York City. *Health Serv Res* 1994; **29:** 95–112.
12. Park RE, Brook RH, Kosecoff J. Explaining variations in hospital death rates: randomness, severity of illness, quality of care. *J Am Med Assoc* 1990; **264:** 484–490.
13. Kahn KL, Keeler EB, Sherwood MJ *et al.* Comparing outcomes of care pre- and post-implementation of the DRG-based prospective payment system. *J Am Med Assoc* 1990; **264:** 1984–1988.
14. Dubois RW, Rogers WH, Moxley JH *et al.* Hospital inpatient mortality: is it a predictor of quality? *N Engl J Med* 1987; **317:** 1674–1680.
15. Dubois RW, Brook RH. Preventable deaths: who, how often, and why? *Ann Intern Med* 1988; **109:** 582–589.
16. Gray D, Hampton JR, Bernstein SJ, Brook RH. Clinical practice: audit of coronary angiography and bypass surgery. *Lancet* 1990; **335:** 1317–1320.
17. Bernstein SJ, Kosecoff J, Gray D *et al.* The appropriateness of the use of cardiovascular procedures: British versus U.S. perspectives. *Int J Technol Assess Health Care* 1993; **9:** 3–10.
18. Pilpel D, Fraser GM, Kosecoff J *et al.* Regional differences in appropriateness of cholecystectomy in a prepaid health insurance system. *Public Health Rev* 1992/93; **20:** 61–74.
19. Brook RH, Park RE, Chassin MR. Predicting the appropriate use of carotid endarterectomy, upper gastrointestinal endoscopy, and coronary angiography. *N Engl J Med* 1990; **323:** 1173–1177.

20. Bernstein SJ, McGlynn EA, Siu AL *et al.* The appropriateness of hysterectomy. A comparison of care in seven health plans. *J Am Med Assoc* 1993; **269:** 2398–2402.

21. Kravitz RL, Laouri M, Kahan JP *et al.* Validity of criteria used for detecting underuse of coronary revascularization. *J Am Med Assoc* 1995; **274:** 632–638.

22. Laouri M, Kravitz RL, Bernstein SJ *et al.* Underuse of coronary angiography: application of a clinical method. *Int J Qual Health Care* 1997; **9:** 5–22.

23. Laouri M, Kravitz RL, French WJ *et al.* Underuse of coronary revascularization procedures: application of a clinical method. *J Am Coll Cardiol* 1997; **29:** 891–897.

24. Froehlich F, Pache I, Burnand B *et al.* Underutilization of upper gastrointestinal endoscopy. *Gastroenterology* 1997; **112:** 690–697.

25. Murata PJ, McGlynn EA, Siu AL *et al.* for the HMO Quality Care Consortium. Quality measures for prenatal care: a comparison of care in six health care plans. *Arch Fam Med* 1994; **3:** 41–49.

26. Hays RD, Brown JA, Spritzer KL *et al.* Member ratings of health care provided by 48 physician groups. *Arch Intern Med* 1998; **158:** 785–790.

27. Brook RH, Williams KN, Avery AD. Quality assurance today and tomorrow: Forecast for the future. *Ann Int Med* 1976; **85:** 809–817.

28. Brook RH. Adapting practice patterns to a managed care environment: carotid endarterectomy – a case example. *J Vasc Surg* 1996; **23:** 913–917.

Accepted for publication 11 April 2000