Robust scientific conclusions are too sparse to inform fully most of the choices that physicians must make about tests and treatments. Instead, ad hoc rules of thumb, or "heuristics," must guide them, and many of these are problematic. Physicians extrapolate from the small samples studied by clinical trials to general populations, but they do so inconsistently. Many physicians live by rules that dictate "not treating the numbers," correcting abnormalities slowly, achieving diagnostic certainty, and operating now to avoid "greater" risk in the future. Yet in each case, historical trends or statistical realities suggest either doing the opposite or investing in more discriminating heuristics. The heuristics of medicine should be discussed, criticized, refined, and then taught. More uniform use of explicit and better heuristics could lead to less practice variation and more efficient medical care.


From the Regenstrief Institute for Health Care, and the Indiana University School of Medicine, Indianapolis, Indiana. For the current author address, see end of text.

Medicine has avoided facing a harsh reality. Although we assume that medical decisions are driven by established scientific fact, even a cursory review of practice patterns shows that they are not. Angioplasty became a multibillion-dollar industry long before a single randomized trial or epidemiologic study had shown its benefits (1). Large numbers of MRI units were purchased during the first 5 years of their clinical availability, before any data were available to show that MRI was preferable to the diagnostic alternatives (2). Also, although randomized trials have proved the lifesaving value of mammography only in women over 50 years of age (3), the American College of Radiology advises that women begin receiving regular testing at 40 years of age.

Scientific data cannot be expected to guide most medical decisions directly. There are not nearly enough randomized trials or epidemiologic studies to guide all clinical decisions (4). In fact, studies showing outcome benefits of testing are almost nonexistent. We have no empirical basis for choosing liver function testing intervals for a patient who is receiving isoniazid or for scheduling electrolyte testing for a patient who is receiving a diuretic. We choose testing and return-visit intervals based on the cycles of the moon—every 3 months or every 6 months—simply because nothing better exists. The benefits of many drug therapies have been established in randomized trials (5, 6). Yet, questions about the best drug to use, the optimal duration of therapy, and the precise severity threshold at which to initiate therapy in a particular patient are seldom answered by these studies.

Only rarely are the sample sizes of scientific studies large enough to specify the ways in which even a few patient factors (for example, age, left main stem artery occlusion) alter the benefits and harms of a therapy. We cannot usually decide the specifics of management for a particular patient directly from reported evidence. However, physicians routinely factor many patient variables (age, comorbidity, social support) into their management decisions.

Physicians clearly do decide, and they generally do so with ease and confidence. Therefore they must have some mechanism, derived from personal theories, assumptions, experience, traditions, and lore, for making these decisions.

Following the lead of Tversky and Kahneman (7), I will refer to the rules of this mechanism as heuristics (8). Cosmologists assume the presence of dark matter because visible matter is not enough to explain the behavior of galaxies. I assume that a set of heuristics explains how physicians make decisions, because available biologic evidence is not sufficient to guide all of their decisions.

Unfortunately, the heuristics used by physicians are poorly understood and rarely discussed. A search of the Abridged Index Medicus for the last 15 years found the word "heuristic" in the title or abstract of only 32 citations. Only four of these citations were original reports about physicians' decision-making processes, and none considered heuristics in general.

It may be that we physicians are reluctant to admit that "nonscientific" mechanisms may guide much of medical care. But admitting the role of heuristics confers no shame. Mathematics—described by many as the "queen of the sciences"—has profitably explored the heuristics that underlie mathematical insight (9). Health care may likewise profit from such an exploration because differences in the heuristics that physicians adopt may explain much of the difference in practice patterns (10). Indeed, such differences do explain much of the
variation in prostatic resection rates among urologists (11).

Exposing these heuristics to critical review so that they can be clarified, improved, and standardized may reduce practice variation, thereby making it easier to optimize the care process (12). Furthermore, we know that many of the "everyday" heuristics described by Tversky and Kahneman (7) are dysfunctional; careful examination of medical heuristics may reveal similar problems and provide corrective insight.

I will identify several specific medical heuristics, examine some of the assumptions that underpin them, and suggest how they might be revised or adjusted on the basis of historical trends, statistical considerations, and appeals to common experience. This discussion should provoke debate about the heuristic rules and sharpen their definition. With serious critique, formal study, and scorekeeping about how well such rules perform over time, we may be able to standardize and improve them. The result should be more reasonable guidance for making decisions about tests and treatment for which decisive empirical studies are not yet available.

Heuristics

A few medical heuristics are familiar to everyone. Occam’s razor is one borrowed from general science (13). Interpreted simply, Occam’s razor advises choosing the simplest hypothesis that explains a set of observations; this heuristic can be directly applied to the diagnostic process. Sutton’s Law is another familiar heuristic, named after the famous bank robber who explained that he robbed banks because “that’s where the money is.” Interpreted in the medical context, Sutton’s Law reminds us to try a common diagnosis to explain symptoms before resorting to an uncommon one, because the statistics favor such bets.

Extrapolation

Extrapolation, the extension of a trend line through known points to another “nearby” point, is justified by the tendency of physical systems to change gradually and smoothly from one point in space–time to another. In medicine, extrapolation is used frequently but not consistently. For example, we extrapolated the outcomes of the Veterans Administration moderate antihypertensive treatment study in men, who were the study participants, to women, who were not represented in the study (5). More importantly, we have also extrapolated the lifesaving benefits of the study drugs (hydrochlorothiazide, reserpine, hydralazine, and propranolol) to every new antihypertensive agent, regardless of its chemical class. Indeed, practitioners have replaced most of the older, tested antihypertensive agents on the basis of this extrapolation. We know that the newer agents reduce blood pressure, but we assume (by extrapolation) that they also save lives. However, recent epidemiologic work indicates that, in the case of calcium channel blockers, this assumption may be wrong (14).

Contrast this acceptance of all antihypertensive agents to our "show me" stance on β-blockers. The β-blocker practolol reduced the death rate from myocardial infarction by one third in a large British study (15). Most of this decrease was due to a reduction in sudden death. Although practolol was never approved in the United States because it caused peritoneal fibrosis, small studies of the β-blockers that were approved in the United States at that time indicated beneficial effects on arrhythmias and trends toward improved mortality (16–18). Yet, no one argued for extrapolating the cardiovascular benefits of practolol to these β-blockers. On the contrary, it was not accepted that β-blockers had cardioprotective effects until large-scale studies of propranolol and metoprolol were completed years later (19). Today we see the same reluctance to extrapolate the lifesaving cardiac benefits of the angiotensin-converting enzyme (ACE) inhibitors, which clinical trials have already shown, to the less expensive, newer ACE inhibitors still under study.

A more consistent heuristic would uniformly adopt the “show me” stance and judge individual drugs solely on the basis of the results of direct clinical trials. This seems to be the current position of the Food and Drug Administration. However, this may not be the best position in a world where information is costly and scarce. A better rule might be to allow extrapolation of therapeutic benefits from one drug to another if they are of the same physiologic class (for example, β-blockers, ACE inhibitors) but not across such classes. Bayesian inference, a statistical method for including all previous information in bets about the future, provides a rationale for exactly this approach (20). The results of completed studies about similar drugs constitute relevant previous information about the efficacy of a new drug. In one sense, it converts all analyses into meta-analyses (21).

History is also supportive of such an approach. Large-scale studies of propranolol and metoprolol showed them to have the same benefit as practolol when they were completed (19), and each of the studied ACE inhibitors has been shown to have the same benefit to congestive heart failure as the first. The validity of extrapolating within classes can be tested as additional clinical trials are done on new species of drugs in one class. By keeping score as the results of studies of H₂ blockers, ACE inhibi-
tors, and HMG-CoA (3-hydroxy-3-methylglutarylcoenzyme A) reductase inhibitors are reported, we can gain further evidence to support or refute this approach.

"Treating the Numbers"

As a resident I was often chided to "treat the patient, not the numbers." I still hear this dictum used to discourage treatment of asymptomatic or mildly asymptomatic patients with abnormal results or findings. Since my housestaff years, most of the arguments about "treating numbers" in the case of asymptomatic patients have been won by the "numbers treaters." In 1970, a Veterans Administration study (5) settled the argument about treatment for mild-to-moderate hypertension in favor of treating patients with diastolic blood pressures greater than 90 mm Hg (22). The arguments about maintaining glucose levels tightly near normal in patients with diabetes have now been settled as well (23). The better the control of glycemia, the better the long-term outcomes. Treatment of asymptomatic diabetic albuminuria (total albumin excretion <500 mg/d) and treatment of asymptomatic hypercholesterolemia in patients with coronary artery disease are now also known to improve patient outcomes (24, 25).

The need to treat the symptomatic extremes of certain measures such as blood pressure (malignant hypertension) and glucoc (diabetic ketoacidosis) may have emboldened physicians to treat abnormalities that do not quite reach the symptomatic extreme. Physicians probably assume that early treatment of the modest abnormalities will keep the patient the same distance from the symptom brink. Experts have argued for decades about whether, and at what threshold, to treat asymptomatic abnormalities. An unassailable position is to treat abnormalities only when they are above the threshold at which clinical trials have shown benefit of treatment. However, we know that the benefits of a new therapy are usually first shown in the patients with the most extreme disorders, because such patient populations have greater statistical power to show the effect of an intervention. From history, we also know that many treatments that initially showed benefits when applied to patients with extreme abnormalities later proved to benefit patients with less extreme abnormalities. Taken together, these two facts argue for choosing treatment thresholds somewhat below the proven boundary when the treatment benefits are great in the most extreme cases and when the moderately extreme cases have not yet been studied. The hordes of physicians who treated diastolic blood pressures between 90 and 115 mm Hg before clinical trial data were available certainly used such an argument.

This is not, however, an appeal for indiscriminate therapeutics. Physicians rarely treat minor abnormalities (less than three or four standard deviations from normal) for any purpose (26). Further, we need to recognize that the benefits of treatment tend to be greater in the more severe cases, whereas the costs and toxicities per patient tend to remain constant. So as we set our treatment threshold lower on the severity curve, the ratio of benefits to costs decreases. The descent in benefits may be rather steep, especially as we get into the thick, less abnormal part of the bell-shaped population curve.

Nevertheless, I stand by the heuristic of "treating numbers." Again, we can test this rule by keeping track of how well it predicts the outcomes of future clinical trials and decision analyses.

Acknowledging Gaps In Our Knowledge of Medicine

The medical community has always been prone to assume that it knows most of what there is to know. This is a common mistake in science. During the 18th century, physicists decided that their real work was done; the only thing left was to grind out the solutions to the Newtonian equations for complicated problems. Then along came quantum mechanics. During the mid-1960s, microbiologists concluded that they had already discovered all pathologic bacteria. Practitioners were harshly criticized for treating pharyngitis not proved to be streptococcal, because they believed that anything else causing a sore throat had to be a virus and was therefore not susceptible to antibiotics. Borrelia burgdorferi, Legionella, Helicobacter pylori, Yersinia enterocolitica, and Twar agent are only a few of the microbes that emerged to show just how wrong these assumptions were. Today, data from DNA sequencing indicate that present culture techniques can grow fewer than 2% of identifiable bacteria, and, in some environments, fewer than 0.01% (27). We should therefore be much more humble in our assumptions about how much we know.

We could speculate that this conceit is an extension of neurologic mechanisms that organize incomplete sensory data into a sensible whole. Just as the brain "papers over" the visual blind spot with the surrounding color and pattern to produce the illusion of a complete image (28), our decision mechanisms suggest that we have a more complete understanding than we actually do. Alternatively, it might simply be a coping mechanism to give us a sense of control where otherwise we might despair. Whatever the origin of this misconception, we should adjust for this tendency to overlook important gaps in our knowledge.

We are also prone to accept too readily practices
based on plausible but unproven assumptions. In the 1960s, everyone "knew" that nitroglycerin should be strictly avoided by patients with myocardial infarction because of the "well known" coronary artery steal phenomenon. Now intravenous nitroglycerin is a routine part of the treatment armamentarium for myocardial infarction. Also, less than 30 years ago, everyone believed that the obligate treatment for hepatitis was long-term bedrest (29). In general, humans tend to have much greater confidence in their knowledge base than is warranted (30).

The problem has two dimensions. The first derives from a habit of thinking in which all current practice is accepted as well founded and correct. We should be much more conscious of the specific rationale for each of our practices and must learn to distinguish those that are based on direct scientific evidence from those that are not. The Evidence-Based Medicine Working Group has enunciated many of the heuristics ("principles") that apply (31), and some reviews now label their practice recommendations on the basis of the specific evidence-based rationale. A recent example is the Fourth American College of Chest Physicians Consensus Conference on Antithrombotic Therapy (32).

However, beneath the lack of critical analysis lurks a second, more pragmatic dimension of the problem. Even if physicians could correct their thinking habits, it would still be too difficult to find and integrate all of the data needed to judge clinical practices. The necessary information is too scattered or is stored in obscure or inaccessible sources. Even when the information can be located, important details may be missing; if the results of a study are negative, the entire study may go unreported.

Therefore, for interventions that lack class A or B evidence of benefit, we should in the meantime adjust our indications and usage rates by some large "humility factor" (33). Once more, I advocate keeping score historically. One should maintain a list of common clinical practices that lack class A or B evidence of benefit—for example, prenatal monitoring or MRI imaging for many indications (34)—and record how often the evidence class levels evolve as they are studied over time.

Forming Realistic Expectations about New Tests and Drugs

In early reports, the distributions of test results for patients with and without a disease often appear to have no areas of overlap, with the test seeming to discriminate perfectly. However, with time and further study, the overlap of the distributions of "disease" and "no disease" becomes apparent. No diagnostic test discriminates perfectly: The diagnosis of pulmonary embolism through arterial blood gas measurement is a perfect example. The first report, based on 40 cases, claimed that a room air PO2 greater than 80 mm Hg reliably excluded the diagnosis of pulmonary embolism (35). The medical community accepted this as law. However, a later report based on a larger sample revealed the overlap: As many as 11% of angiographically proven pulmonary embolism cases will have a room air PO2 greater than 80 mm Hg (36). Other examples, such as the use of the ACE test to diagnose sarcoidosis and the use of carcinoembryonic antigen to detect cancer, can easily be found (37, 38).

Jaeschke and colleagues (39) provide several specific criteria for judging reports of test efficacy that would help physicians avoid such errors. However, only with large and repeated studies can we be certain of the true discriminating power of a test. Individual practitioners should discount early claims of great performance by some large "doubt" factor. Journal editors should also apply stronger editorial control to eliminate design flaws in reports about the operating characteristics of tests.

An analogous problem occurs with drugs. New drugs often become the "preferred" drugs in their class within a year or two of their introduction. It is as though we apply a rule that newer is automatically better. Burnum (40) describes the phenomenon thus: "Giving a new drug spreads as a contagion from one physician... to another." When drugs have unique actions, this may be justified. Here the problem is that physicians fail to recognize the risks lurking beneath premarking (phase 3) safety data. Such data are only "guaranteed" to reveal acute toxicities that occur more often than 1 per 100 administrations. Toxicities that occur less often than 1 per 1000 administrations or that take more than 6 months to evolve will not be revealed (41). This is not a sufficient margin for replacing an older drug for which the unknown risks are much lower. Remember that it took less than 1 case of aplastic anemia per 20,000 patients who had been treated to make chloramphenicol a drug of last resort.

In this light, it is not surprising that many drugs have revealed their "dark" side after they have become popular. Clindamycin rapidly replaced cefazolin as the preferred gram-positive antibiotic for potentially septic patients in the early 1970s. Reports of pseudomembranous colitis in as many as 10% of patients treated with clindamycin reversed this trend (42). Between 1982 and 1983, zomepirac, a nonnarcotic said to have a strength equivalent to that of morphine became the most popular prescription medication for pain. Reports of anaphylaxis occurring 10 to 30 times more frequently than with other nonsteroidal drugs and more than 30 deaths chased it from the market in 1983 (43). Ticrynafen, introduced in 1979, was the first major diuretic that did
not raise serum uric acid levels (44). It never became the most popular diuretic, but it did have a large market share when it was withdrawn 1 year after its introduction because of deaths due to liver failure. Triazolam has rapidly become the most used hypnotic agent in the United States, despite the fact that its incidence of rebound insomnia and anxiety is 16 times greater than that of its competitors (45). Great Britain has withdrawn its approval of this drug.

Such cases suggest the need for a new rule about new drugs: “Never use a new drug when an old drug will do.” With this rule, new drugs would be prescribed only for their unique indications and for patients who cannot tolerate the available older drugs. Such an approach would give us time to discover the hidden dangers of new drugs and might reduce drug costs, as well.

These examples all illustrate the problems that may arise when technology is adopted too rapidly. However, the opposite problem also occurs. Fewer than 25% of eligible patients receive influenza vaccination each year, despite the fact that it has been several decades since its introduction (46). Far too few diabetic patients have their recommended eye examinations, even though randomized trials have proved the benefit of such examinations to sight (47). Clearly, the degree to which a product is actively marketed and the immediacy of its benefit have an influence on whether an agent is used. Pharmaceutical manufacturers could probably teach us many things about correcting such problems of underutilization.

Avoiding Narrow Definitions of New Syndromes

We tend to define new disease syndromes too narrowly. We identify a disease initially on the basis of the most severe cases and the signs and symptoms that are easiest to observe. However, time almost always reveals a broader spectrum of presentations and milder forms. For example, asthma, which is classically defined in terms of bronchospasm, has a variant that presents as cough alone (48); toxic shock may occur without shock (49); and cardiac angina may occur without chest pain. In fact, some authorities now believe that most episodes of angina are actually painless (50). Yet, practitioners are very slow to modify the first, rigid definition of a disease. Therefore, I propose another clinical heuristic for consideration: Avoid rigid definition of disease states and assume that all disorders have many variants. One corollary of this heuristic might be a willingness to ignore at least one of the defining criteria when trying to match a patient’s findings to alternative disease templates.

In the case of tests, drugs, and diagnoses, we should adjust our initial perspectives on the basis of statistical realities. All initial perspectives should be considered tentative and incomplete. Only with greater experience (and larger sample sizes) will the picture become clear.

Keeping Treatment from Becoming Subservient to Diagnosis

Osler beat one dictum into the heads of his students and all internists who followed. The three steps of patient care are: diagnose, diagnose, diagnose. Because action often preceded thought in Osler’s day, this edict was probably needed to counterbalance thoughtless empiricism. However, Osler’s rule may have pushed the pendulum too far, and as a result, internists may now overtest and undertreat (51). Reports document undertreatment of pain (52) and depression (53). Internal medicine house officers dog the laboratory for the results they ordered, yet rarely invest the same energy in making certain that prescribed treatments have been started (54).

A diagnosis is not a prerequisite to treatment. In the case of a disease with one treatment and one test option, Pauker and Kassirer (55) proved that when the prior probability of the disease exceeds the treatment threshold, the right choice is to treat without testing. For sore throats, their equation translates into “treat when the probability of streptococcal infection is greater than 20%” (56). Most experienced practitioners do not hesitate to initiate a therapeutic trial when the patient is asymptomatic and the diagnostic state is unclear. However, empirical therapy is often disdained in academic medical centers, perhaps because the “placebo effect” makes such trials difficult to interpret. But much of the “placebo effect” in the routine prescribing of oral medication can be explained by statistical regression, the effect of which can be avoided (26). Alternatively, we could use “n-of-1” trials to further clarify the interpretations of therapeutic trials (57). In any case, we need to study and formalize the rules about the use of therapeutic trials so that we can convey practice realities to our trainees.

Surgical Procedures in Asymptomatic Patients

Two assumptions underpin arguments for elective surgery when symptoms are mild or absent. The first is that the clinical problem targeted by the surgery will worsen inextricably over time, so that the patient will require the surgery eventually. The second is that delayed surgery will be more difficult or dangerous because of emergency circumstances or increased age.

Data suggest that differences in the rate of transurethral resection among urologists can be explained by differences in their acceptance of these assumptions about surgery (58). However, careful analysis suggests that these assumptions are often
wrong and that watchful waiting is a better option than elective surgery. Decision analysis shows that watchful waiting has fewer life risks than early prostate surgery when the patient has mild-to-moderate retention (<200 mL postvoid) (59). Watchful waiting is also a better choice than prostatectomy for many categories of patients with localized prostate cancer (60), and it is a better choice than open surgery for asymptomatic gallstones (61). Careful risk-benefit analysis may reveal the advantage of this approach for many elective procedures.

Other considerations weigh on the side of watchful waiting. The passage of time offers the chance of technical breakthroughs (for example, laparoscopic cholecystectomy instead of full laparotomy, and lithotripsy instead of kidney surgery). Also, community-based estimates of perioperative death rates, taken from Medicare or regional databases or both, are usually much higher than the death rates that are commonly quoted in the medical literature (62). The favorable rates usually come from highly selected patient populations at academic medical centers and therefore may not apply to the community at large. The net result may be that procedures are done far more often than would be justified if the higher community mortality rates were factored into the decision.

When randomized trials show no clear benefit of so-called preventive surgery (63), it may be better to follow a new rule of thumb for certain situations: Don't operate when the patient can wait.

Some surgical procedures, on the other hand, may be underutilized. Knee replacement surgery may be such a case, considering the functional benefits that accrue to patients and the low rate of its use in some parts of the United States (Heck D, Personal communication). But with the abundance of good surgeons and positive financial incentives in the United States, most reported concerns about surgical procedures are about overutilization.

Conclusion

General rules for weighing scientific evidence have been formally presented in the medical literature for several years (64). These rules were not derived from formal analyses of empirical data. They were derived from common sense or from some other fundamental principles. That is how all logical processes begin, whether in mathematics or medicine. First principles, axioms, postulates, and rules of thumb—what we have lumped together as heuristics—always provide the starting point. Making decisions about the day-to-day care of patients should not be any different. Yet, medicine almost ignores (or perhaps denies) the role of heuristics in medical decision making.

In this report, I have described and critiqued some of the heuristics seen in day-to-day medical practice. These heuristics should be judged on the basis of historical experience (how well the heuristics have predicted the outcome of clinical trials), statistical principles (small samples do not accurately represent general populations), comparisons of underpinning theories with real data (for example, the assumption that early surgery is better than later surgery), and how consistently a particular heuristic is used.

Because of ethical and cost considerations, there will never be enough randomized trials or epidemiologic studies to guide every clinical decision; heuristics provide personal criteria for making such decisions. Because physicians will always need to depend on them to some extent, the medical community should begin to consider these heuristics more closely—to identify, criticize, improve, and standardize them. We should be more rigorous in identifying the assumptions, rationale, and evidence on which our current practices are based. If we cannot justify a decision on the basis of clinical trial data or a robust heuristic, perhaps pure cost factors should control the decision. As randomized trials and decision analyses are completed over time, we should also keep score. The heuristics that hold up—by being confirmed by the results of future studies—should be retained and re-enforced. Those that do not should be discarded.

From the beginning, the Canadian Preventive Care Task Force has been explicit about the rules it uses to judge evidence for a preventive intervention (65). The disclosure of these rules and the Task Force’s consistent adherence to them adds a simple elegance and unity to their conclusions. More recently, Canadian investigators have defined rules for evidence-based medicine (31). We need to apply the same systematics and formalism to the many and varied rules in the many shadows where scientific evidence does not give clear direction. We need to list discrete heuristic rules; to estimate, based on statistical realities and past experience, which ones are best under which circumstances; and to encourage providers to standardize a set of heuristics. Finally, we must keep score: Which rules consistently predict the outcome of clinical trials when such formal studies are finally available.

Grant Support: By grants HS 07719-013 and HS 08750-02 from the Agency for Health Care Policy and Research; and by grant N01LM-4-3410 from the National Library of Medicine.

Requests for Reprints: Clement J. McDonald, MD, Regenstrief Institute, 1001 West 10th Street, Indianapolis, IN 46202.