Making Policy When The Evidence Is In Dispute

Good health policy making involves consideration of much more than clinical evidence.

by David Atkins, Joanna Siegel, and Jean Slutsky

ABSTRACT: Policymakers often struggle with medical issues that are the subject of fierce scientific debate. On closer examination, many of these debates are manifestations of conflicting perspectives and values as much as disagreements over the evidence. We summarize common factors underlying recent debates and outline a series of questions that can help disentangle questions of evidence from those of values. These questions focus on identifying the most important outcomes, evaluating the quality of evidence, and assessing the trade-offs involved. We then use four recent policy debates—involving prostate-specific antigen (PSA) screening, high-dose chemotherapy for breast cancer, antibiotic therapy for otitis media, and newborn hearing screening—to illustrate how this approach can help clarify areas of agreement and disagreement of the opposing sides.
choose sides in complicated scientific arguments.²

Why haven’t evidence-based principles, which were intended to replace the subjective opinions of experts with an objective, explicit analysis of science, reduced the clashes surrounding policy decisions? We propose that many of the debates that appear to be intractable disputes over the evidence arise from conflicts in the other spheres that influence decisions, such as the values, preferences, and circumstances of individuals and the communities they represent. When the evidence itself is a subject of debate, we argue that a formal evidence-based approach can help separate questions of evidence from the other important considerations. This process often reveals surprising consensus on the scientific evidence, which is masked by fundamental differences of opinion about what outcomes are most important and what actions are appropriate in the face of imperfect evidence.

We outline below a series of questions that can help policymakers disentangle questions of evidence from those of values and preferences and weigh the credibility of competing arguments about evidence. These questions focus on identifying the key aims of the policy, evaluating the quality of evidence, and assessing explicitly the trade-offs involved. We then use several recent clinical policy debates to illustrate how this approach can help clarify areas of agreement and disagreement and help decisionmakers understand the perspectives of the opposing sides.

A Framework For Evaluating Debates About Evidence: Questions For Policymakers

What is the ultimate goal, and how does the intervention achieve those ends? Heated debates can sometimes obscure the ultimate aim of a policy decision and the relationship of the policy intervention to that outcome. Explicitly examining the steps that link the policy intervention to the most important outcome serves three important purposes. It identifies distinct questions to which the evidence can be applied. It makes explicit the fact that policy decisions have a range of intended and unintended outcomes. And it helps distinguish when the evidence directly addresses the important effects of policy from when the effects must be inferred from less direct evidence.

Teasing out embedded issues. Schematics are a useful tool for mapping available evidence to key policy questions and outcomes.³ A policy question, such as whether the new Medicare drug benefit will reduce the costs of treating chronic disease, often has imbedded in it a series of separate clinical or policy issues: How much will Medicare coverage actually reduce out-of-pocket costs for drugs to treat chronic diseases such as diabetes and heart disease? How much will reduced costs increase compliance with drug regimens? How will this change in compliance improve intermediate outcomes such as blood pressure, lipid levels, and blood glucose? How will this translate into better outcomes (fewer heart attacks, stroke, and so forth) and the costs of treating them? How do the costs of drug coverage compare with alternative strategies for promoting compliance (for example, pro-
moting generic drugs, improving clinician decisions, patient education)?

Sometimes, available evidence addresses the effects of policy changes directly. For example, when the New Jersey Medicaid program imposed a cap on drug coverage, researchers were able to show a resulting increase in nursing home admissions. More often, policymakers must examine evidence addressing component questions to elucidate the ways in which a policy measure will affect health.

Relative importance of outcomes. Debates over evidence may reflect disagreement over the importance of different outcomes. Intermediate (or surrogate) outcomes mediate the effects of the intervention on more important health outcomes, but they may be valuable in their own right to some groups. Advocates may value an educational campaign shown to increase knowledge about HIV (an intermediate outcome), but policymakers may want evidence that it reduces the number of new HIV infections (a health outcome). Different perspectives may even shape what constitutes a health outcome. To obese patients, weight loss itself may seem like a health benefit. Medicare, however, will cover only those obesity treatments that improve health outcomes such as heart disease risk or blood pressure.

Identifying target populations. Identifying the target population is important in specifying a policy question. There may be consensus on the value of an intervention in one group (such as tuberculosis testing in new immigrants) but disagreement about extending the policy to other groups (such as annual testing of all schoolchildren). Disagreements may also arise if the intervention is evaluated with respect to different alternatives. The benefits of providing fluoride treatments through Medicaid will differ if it is compared with fluoridation of public water versus no treatment at all.

How good is the evidence that the intervention can improve important outcomes? Policymakers need not be expert at evaluating evidence, and there is no single approach to grading evidence for the variety of issues confronting them. Policymakers can go a long way toward understanding the scientific basis of many debates if they keep in mind a hierarchy of questions suggested by Brian Haynes for evaluating any health intervention: Can it work? Will it work? Is it worth it? Conflict can arise because evidence to address the first question may not answer the other questions, which may be most important to some decisionmakers.

Determining whether something can work requires studies of high validity (usually termed “internal validity” by epidemiologists). A valid study protects against bias and produces findings that are likely to be true. Well-conducted randomized controlled trials (RCTs) provide the best protection against bias, which is why they are the standard for some policies such as drug approval, but they are difficult or impossible to conduct for many policy questions. Policy researchers usually must use observational (that is, nonexperimental) study designs, which are more vulnerable to bias. This can be reduced, however, by choosing appropriate control groups, using representative study populations, and statistically adjusting for other factors (confounders) that might explain the effect.
How good is the evidence that the intervention will work in my setting?
Policymakers must next consider whether a study’s results are relevant or applicable to the question of interest (usually referred to as “external validity”). The effect of an intervention in one setting or population may not translate to another setting. Moreover, an intervention’s performance under ideal conditions (efficacy) can differ dramatically in actual practice (effectiveness). There is an inevitable tension between the high internal validity provided by careful RCTs and the realization that such trials may not give an accurate picture of the impact of a policy decision under real-world conditions.

An intervention is more likely to work in practice when evidence can link it directly to important health outcomes rather than intermediate outcomes. Parties often differ on the appropriateness of relying on intermediate outcomes, be they related to treatments (for example, tumor size) or policy (for example, proportion of children receiving drug abuse education). Intermediate outcomes have led clinicians astray in notable cases, resulting in widespread use of ineffective or harmful medications such as anti-arrhythmic drugs and postmenopausal hormone replacement therapy (HRT), but other intermediate outcomes (for example, smoking quit rates) are widely accepted as reliable predictors of health outcomes.10

Finally, results that are consistent across studies provide reassurance that findings are not the result of chance or particular to one community or institution. Often, promising initial results are not borne out by future studies.11 In rare cases, studies offer such clear results that they cannot be replicated because of ethical concerns; in general, however, policymakers should be skeptical of evidence derived from a single study. A central tenet of evidence-based decisions is the importance of comprehensive, objective reviews that use all available evidence, published and otherwise.12 The realization that the published literature may not always give a complete picture of the true effectiveness of drugs has led to recent calls for registries of all pharmaceutical trials.13

How do the potential benefits compare with the possible harms or costs of the intervention? All important decisions involve trade-offs. Deciding whether a policy is “worth it” requires considering the outcomes affected by the policy, the size of individual benefits and harms, the degree of uncertainty around these estimates, and the importance (values) of the different outcomes. Judgments about trade-offs are inherently expressions of values. Patients and specialists highly value possible benefits for the conditions they have or treat, even if the benefits are potential rather than proven and even if they represent marginal rather than major advances. Decisionmakers in health plans and public health agencies worry about broader impacts to the health system, including economic and opportunity costs, harms, and effects on access and equity—critical issues for society that rarely resonate with individual patients. Two clinical trials reported that teaching breast self-examination did not reduce cancer deaths, but the possibility that it might help even a few women provided sufficient justification to advocates to continue promoting self-
examination, which also helps promote breast cancer awareness. To primary care clinicians, the time diverted from more important issues and the false alarms and unnecessary biopsies for many healthy women are compelling reasons to abandon teaching breast self-examination until there is clearer proof of benefit.

Although a frequent source of conflict, cost issues are often submerged in evidence debates. Compelling stories of children who died from very rare metabolic disorders that might have been detected with newer, more expensive equipment have created powerful momentum for expanded screening of newborns. But in an era of constrained budgets, state policymakers need to weigh the benefits and costs of new screening programs against those of other equally important programs. Nonetheless, it remains politically risky to frame a health policy decision as being based primarily on cost or cost-effectiveness.

What constitutes “good enough” evidence for a policy decision? The judgment about what degree of certainty is sufficient is inherently a judgment about the risks of acting too soon (promoting a policy that turns out to be ineffective or harmful) and acting too late (delaying a beneficial intervention to get better evidence). Controversy is most common in areas where definitive studies are lacking or inherently difficult. When the Environmental Protection Agency (EPA) concluded that environmental tobacco smoke caused lung cancer, it provoked controversy by relying on a meta-analysis of epidemiologic studies and by relaxing the usual standard for statistical significance. Whereas the usual threshold for defining a significant result is $p < .05$, the EPA decided that $p < .10$ (equivalent to less than a 10 percent probability that the observed link between environmental tobacco smoke and lung cancer was due to chance) was sufficient for policy making. To public health policymakers, considering the bulk of supporting evidence, the secondary benefits of smoke-free workplaces, and the risks of waiting for more definitive studies, the evidence was “good enough,” an opinion understandably not shared by the tobacco industry.

Acting too soon carries real risks, however. Promoting unproven interventions can waste money and harm people, sometimes substantially. The strict standards for Food and Drug Administration (FDA) drug approval generally reflect the high value Americans place on having medications that are proven effective and safe. To policymakers, acting too soon carries another important risk: Early dissemination of a new medical intervention may make it impossible to conduct definitive studies. Acting too soon also can make it difficult to assess other options or change course. These tensions were played out in the debate over the early adoption of helical computed tomography (CT) for lung cancer screening. Once CT was shown to be better than chest x-rays for detecting early lung cancer, proponents argued that it should be disseminated while ongoing observational studies examined whether it reduced mortality. The National Cancer Institute (NCI), concerned that this costly technology might become entrenched in practice before the benefits and potential harms (frequent false-positive results and possible harms from surgical procedures and other treatment) were adequately tested, argued success-
fully for a more rigorous trial.18

■ What other considerations are relevant to policy decisions? Often, policymakers face considerations that are not captured in a neat accounting of benefits and harms.19 Concerns about equity may have implications for the targeting of policies: Equity may favor emphasizing interventions with potential benefits for special populations at risk or, conversely, may favor interventions that can reach the whole population over those that would be available to only a few.20 Policymakers in the private sector may need to consider the impact of decisions on their market—health plans may feel compelled to cover some new unproven interventions because they are valued by their members, because it helps their marketing, or because denying them will cause adverse publicity or potential liability issues.

Four Recent Clinical Policy Controversies

We use four recent clinical policy controversies to illustrate how debates about the evidence have arisen from predictable differences in the perspectives of opposing parties. Policymakers who understand this dynamic can then make decisions by judging which perspective most closely reflects their responsibility to their constituents, without feeling that they have to be the final arbiters of truth.

■ PSA screening for prostate cancer. The ongoing debate about screening for prostate cancer with prostate-specific antigen (PSA) illustrates how our perceptions of trade-offs are shaped by our perspectives. The effectiveness of PSA screening to reduce prostate cancer deaths has remained controversial since the first reports emerged that it could identify early prostate cancer. Despite this controversy, PSA screening has become increasingly common: In recent statistics, more than 70 percent of men over age fifty have been screened at least once.21 Congress and state legislatures confronted this issue in the 1997 decision to extend Medicare coverage and state mandates to require insurance coverage of PSA testing.22 Although the American Cancer Society promotes offering screening to men older than age fifty, many primary care societies have refrained from recommending routine screening.23 This split conceals a fair consensus on the basic findings: PSA testing can detect early prostate cancer but produces frequent false-positive results, especially in older men; some cancers detected by PSA are slow growing and do not need to be treated, especially in men over age seventy; no studies have definitively demonstrated a reduction in prostate cancer deaths attributable to screening; and prostate cancer mortality in the United States has been falling recently. Nevertheless, the debate continues.

Specialty versus primary care groups. The contrasting positions of specialists and primary care groups reflect differing views of PSA screening’s potential benefits and harms. To cancer specialists, cancer organizations, and an increasing proportion of the public, early detection is a worthwhile end in itself.24 Independent of effects on mortality, early detection provides important information, reduces the regret that patients diagnosed with advanced cancer often experience, and may provide patients with more treatment options. Negative screening results can re-
assure anxious patients. From the vantage point of clinicians and advocates, the downsides of false-positive results and potential overtreatment (treating slow-growing cancers that might not have been discovered) are unavoidable aspects of cancer screening and an acceptable price to be paid for the hope of saving lives. Finally, waiting years for definitive trials while there are no other good options for reducing prostate cancer deaths is judged an unacceptable option.

Nonspecialty organizations and public health decisionmakers in state cancer control programs, in contrast, view the issue of PSA screening through a prism of primary care practice and public health and the wide array of patients and prevention issues for which they are responsible. The possibility that screening might actually do more harm than good is a much more real concern, especially since tangible harm—including incontinence, impotence, and death from surgery—may occur in previously healthy men. Given the difficulty of adequately delivering proven preventive care, adding on new, unproven services has sizable opportunity costs. Finally, primary care physicians know that the option of informing patients about the pros and cons of screening and letting them share in decision making is difficult in practice. Even informed patients may experience great anxiety from false-positive tests, and once diagnosed with cancer, few patients are willing to consider the option of doing nothing (“watchful waiting”) even when the risks of treatment might exceed the risk from a slow-growing cancer. Despite agreement among many experts that older men have the least to gain from screening and treatment, this group now accounts for the majority of prostate cancer screening and treatment.

Concerns other than evidence. Concerns other than evidence have also clearly affected policy decisions on prostate cancer. Many legislators have been diagnosed with prostate cancer. Supporting screening has been a way for others to indicate their support for men’s health issues. With many men availing themselves of screening, mandating insurance coverage has been seen as a way to ensure that access to screening did not depend on ability to pay. Insurance coverage, however, helped further the impression that screening had become the standard of care; indeed, successful suits have been brought against clinicians who have not actively recommended screening.

■ High-dose chemotherapy and bone marrow transplant for breast cancer. The saga of high-dose chemotherapy and autologous bone marrow transplant (HDC-ABMT) for aggressive breast cancer is a sobering case study of the challenges of deciding when emerging evidence is sufficient to make policy decisions in the face of intense public scrutiny. Faced with the grim prognosis of women with advanced breast cancer, clinicians began experimenting with higher doses of chemotherapy, using ABMT to rescue patients from the toxic effects. Based only on limited evidence from Phase II trials, with no controls or only historical controls, experts were soon touting this new treatment and clashing in courts with those who questioned its benefits. Hospitals rapidly added this to their treatment alternatives, and law-
suits and legislative actions pressured insurance companies to cover the treatment, which could exceed $100,000 per case. Only after government agencies and insurers collaborated on a series of definitive trials—and after more than 20,000 patients had been treated at a cost of more than $2 billion and an estimated 600 premature deaths had occurred—was the procedure demonstrated to provide no benefit (and a higher risk of early death) over conventional therapy.28

Acting on premature evidence. The ABMT example illustrates that what one considers "sufficient" evidence for a policy decision is dramatically different if one is a patient responsible to oneself rather than a policymaker responsible to the larger population. To patients and their treating physicians, the risk of acting too late—forsaking a potentially superior treatment—loomed much larger than the risk of acting too soon. This translated into a willingness to rely on favorable results of early studies using intermediate outcomes (shrinkage of the tumor), even as definitive trials were ongoing to test whether it really improved survival. Facing a grim prognosis, some patients may have felt that they had little to lose from a new but unproven treatment. Less attention was given to the sizable risks of the treatment and its adverse effects on quality of life.

Considerations other than evidence. As in PSA screening, considerations other than evidence influenced practice. The high costs of treatment made it a profit maker for hospitals and elevated the position of its proponents. State policymakers were understandably reluctant to side with insurers and against patients, and they could console themselves that mandating coverage merely allowed greater latitude for patients and their physicians to choose the “best” treatment.

To those responsible to the larger population of cancer patients (including future patients) and the general public, the risks of acting “too soon” looked much greater. Researchers, policymakers for health plans, and groups such as the National Breast Cancer Coalition (NBCC), all of whom had a commitment to evidence-based policy, recognized the need for carefully conducted trials with clinical endpoints. As the NBCC noted after the initial reports of disappointing results in the clinical trials, “Unfortunately, because so many physicians performed this procedure outside of a clinical trial setting, we do not know how effective it is. Had these procedures been performed within a randomized clinical trial we would have had the answers some time ago.”29 The high cost of ineffective therapy was borne by a vulnerable group of patients and the public at large, while the role of politicians mandating individual treatments set a troubling precedent in an era of high-tech medicine and vocal interest groups.

Inappropriate antibiotic use. Occasionally new outcomes emerge that change the substance of the policy debate. Until recently, clinical practice and clini-
cal practice guidelines in the United States promoted antibiotics for uncomplicated middle-ear infections (otitis media) in children. Practice in countries such as the Netherlands have for some time promoted more conservative therapy with analgesics alone, but the routine use of antibiotics has been reexamined by U.S. organizations only recently as the effects of antibiotic use on antibiotic resistance have become clearer. The basic evidence on the benefits of antibiotic treatment for uncomplicated otitis media has not changed greatly: The large majority of children improve within twenty-four to forty-eight hours and symptoms resolve within a week without antibiotics. Antibiotics modestly increase the speed at which symptoms resolve and reduce the numbers who fail to get better, but between eight and twenty children need to be treated to provide modest benefits to one. At the same time, one in six children may have antibiotic-induced vomiting or diarrhea.

Why is treatment so common? From the perspective of many practitioners, the small benefits exceeded the risks, especially when considering that parents had come to expect antibiotics for their sick children. Policymakers, including the Centers for Disease Control and Prevention (CDC) and primary care professional societies, have recognized that the balance of benefits and harms for society differ from those for the individual patient. Antibiotic resistance is increasing because of overuse of antibiotics, not only for ear infections but also for viral illnesses. Rising resistance is reducing the effectiveness of newer drugs in treating life-threatening illnesses. The benefits of routine antibiotics look even smaller if compared with the strategy of reserving antibiotics for those children who are not improved after observation in the first two days.

Reframing the benefits and risks has had an effect. A recent study of antibiotic use in children during 1996–2000 showed a dramatic decrease in use. Not only did antibiotic use decrease, but the proportion of “first line” antibiotics prescribed increased, which indicates reduced use of newer broad-spectrum antibiotics.

**Newborn hearing screening.** Policymakers in many states have had to wrestle with debates over newborn hearing screening, which is mandated in thirty-eight states. When the U.S. Preventive Services Task Force (USPSTF) concluded in 2001 that the evidence was “insufficient to recommend for or against universal screening,” advocates accused it of ignoring the recommendations of many other organizations and setting an unrealistic standard for proof. The important policy debate obscured what was actually fairly widespread agreement over the scientific evidence. The first two years are a critical period of language development, and infants whose hearing problems go undetected have delayed language development. Universal newborn hearing screening can detect most infants with serious hearing loss. Initial tests generate a sizable number of false-positive results that require further testing, but serious harms from testing have not been documented. Selective screening of high-risk infants has a higher yield but will miss many deaf children. Screening leads to earlier diagnosis and treatment—earlier on average by six to twelve months—but studies of the long-
One source of disagreement is which outcomes are of primary interest. The idea that late diagnosis produces lasting speech and language problems has been a major driver in policy initiatives to expand screening, although the USPSTF found the evidence for this to be relatively weak. To advocates, however, detecting hearing impairment provides a number of other important benefits: One can inform parents about a child’s disability, prevent the frustration and guilt many parents feel if a serious hearing problem is picked up late, and prevent the rare cases of children going many years with undetected hearing problems.

The potential benefits of universal newborn hearing screening also depend on the alternatives considered and the assumptions about how well it will work in practice. Proponents typically depict the benefits of universal screening compared with no screening, but benefits are much smaller compared with the current standard of care, which includes screening high-risk infants. While the USPSTF noted that problems in follow-up, diagnosis, and treatment will further reduce the real-world benefits of screening, proponents see universal screening as an important step to advocating for better resources where they do not already exist. There are also issues of equity: Universal screening is the only way to ensure an equal chance of detecting all hearing-impaired children, regardless of where they are born or the quality of their medical care after birth. Ultimately, parties differ most strongly over the need for better evidence. Advocates accused the USPSTF of setting an unrealistic standard for practice and requiring studies that would be difficult and perhaps not ethical, given the prevalence of screening and treatment.

How then should a policymaker negotiate these competing perspectives? As described above, questions of values and resources are central to the debate. The most difficult policy question is gauging the implications of investing new resources in screening. If expanded screening can be matched by the necessary resources for expanded follow-up and treatment—resources that need to be organized at the state level—policymakers might decide that it is appropriate despite some uncertainty. But if resources are directed only toward screening or come from diverting resources from other worthwhile programs for infants and children, then the trade-offs are much less clear. Having better data on the clinical and economic value of hearing screening becomes critical.

An evidence-based approach to health care policy decisions will neither eliminate controversy nor relieve policymakers of the difficult task of making decisions in the absence of clear scientific consensus. A systematic consideration of the questions outlined above, however, can help disentangle the separate factors that contribute to policy disputes. It can allow decisionmakers to identify specific questions at the heart of a policy issue; evaluate the evidence available to answer those questions and define important areas of uncertainty; and outline the trade-offs between the most likely benefits, harms, and costs. Differing
values and resource constraints can produce conflict even when there is good evidence and the policy outcomes are clear. Conflict is most common, however, when evidence is weaker, outcomes are less certain, and parties disagree about the risks of acting in the face of uncertainty. Being explicit about these elements of decision making is as important as an “evidence-based” approach in ensuring that decisions are transparent and consistent with both the science and the values of individuals and society.

The authors are grateful to Susan Norris for her thoughtful comments on the manuscript. The opinions expressed herein are those of the authors alone and do not represent the official positions or policies of the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

NOTES


