

**Being Smarter About Clinical Trials:
A Report of the NIH Workshop**

***Moving From Observational Studies to Clinical Trials:
Why Do We Sometimes Get It Wrong?***

held January 11–12, 2005

Barnett S. Kramer, M.D., M.P.H.; Joan Wilentz, M.A.; Duane Alexander, M.D.;
John Burklow; Lawrence M. Friedman, M.D.; Richard Hodes, M.D.;
Ruth Kirschstein, M.D.; Amy Patterson, M.D.;
Griffin Rodgers, M.D.; Stephen E. Straus, M.D.

All opinions in this paper represent those of the authors and meeting participants and do not necessarily represent official views or positions of the Federal Government or the Department of Health and Human Services.

Abstract. Unexpected outcomes, including harms, of several recent large clinical trials, both publicly and privately sponsored, prompted a meeting at the National Institutes of Health (NIH), on January 11–12, 2005, entitled *Moving From Observational Studies to Clinical Trials: Why Do We Sometimes Get It Wrong?* Speakers analyzed the flaws in a number of trials and other clinical research studies. These flaws included traditional sources of error, such as systematic bias as well as insufficiently validated biomarkers and surrogate endpoints. Participants addressed ways to maximize the quality of evidence available by using traditional statistical tools as well as by adopting methods and frameworks by which to validate biomarkers and surrogates and enhance the value of meta-analyses and systematic reviews. They also considered the impact of the “-omics” revolution on the design and statistical methodology to be used in future clinical trials. The importance of accurate and timely reporting of trial outcomes was highlighted in discussions of how trials are (or are not) reported in the literature, in the media, and in medical advertising. An underlying theme was the extent to which investigators must be continually on guard against conventional “wisdom,” intuitions, and unwarranted assumptions that can lead to conflating causal and noncausal associations and other sources of error. An underlying theme of the conference was Aristotle’s observation in *Metaphysics II*: “The investigation of truth is in one way hard, in another way easy.” The authors used the group’s discussions to generate a set of recommendations for improving the quality of clinical investigations, including criteria for setting priorities and for moving trial results expeditiously into health policy and decisionmaking.

In July 2002, the NIH halted the Women’s Health Initiative (WHI) clinical trial of combined estrogen plus progestin hormone replacement therapy (HRT). Statistical analysis indicated evidence of increased risks of breast cancer, heart disease, and stroke among postmenopausal women taking the combined estrogen-progestin regimen.^{1,2} In September 2004, Merck withdrew its cyclooxygenase-2 (COX-2) inhibitor rofecoxib (Vioxx™) from the market because a clinical trial showed increased risk of heart disease and stroke among long-term users of the pain-relieving drug.^{3,4} Concerns about rofecoxib and other COX-2 inhibitors prompted the Food and Drug Administration (FDA) to seek an expert committee’s advice on the drugs’ safety. In the interim, a randomized clinical trial of celecoxib (Celebrex™) for prevention of recurrence of colorectal polyps was put on hold.⁵ The FDA committee has since met and advised that the drugs Celebrex and Bextra™ (valdecoxib) made by Pfizer could remain on the market, but they should not be advertised directly to consumers and must carry stringent “black box” warning labels about cardiovascular risks.⁶ At the time of the meeting, the fate of Vioxx was yet to be decided. In its most recent decision, FDA has now asked Pfizer to withdraw Bextra from the market.⁷ If those front-page stories are combined with the report of an athlete’s death attributed to the dietary supplement ephedra⁸ and the recent FDA advisory that selective serotonin reuptake inhibitor (SSRI) drugs used to treat depression may pose a suicide risk in adolescents,⁹ there would seem to be cause for public health concern about the scientific process that brings medical interventions to the market. How could these serious harms have come about despite earlier studies that suggested favorable benefit-to-risk ratios and appeared to ratify conventional wisdom? Mindful of the need for thoughtful analysis, Elias A. Zerhouni, M.D., Director of NIH, initiated a meeting on January 11–12, 2005, that brought clinical investigators, trialists, biostatisticians, and

other experts together with NIH scientists to discuss *Moving From Observational Studies to Clinical Trials: Why Do We Sometimes Get It Wrong?*

“It is time for an ‘M and M’ [morbidity and mortality] conference [on medical evidence],” Dr. Zerhouni said at the opening of the meeting. “Are we really challenging ourselves to use better tools, better methodologies? Could we have come up with better surveillance or study designs to have alerted us earlier [to unexpected results in these clinical trials]?” He challenged the group to come up with innovative ideas and frameworks and to exploit new technologies to aid in making decisions and policies, as the credibility of the scientific enterprise was at stake. “Forty percent of science news relates to health or medicine,” he noted, “and we are seeing a gradual erosion of public trust.”

In the case of the WHI trial, conventional wisdom—based primarily on nonrandomized observational studies—so strongly supported the view that HRT would *lower* the risk of heart disease, stroke, and dementia in postmenopausal women, as well as prevent hip fracture, that when plans for the study were proposed in the early 1990s, critics protested that NIH was wasting money on a Phase 3 clinical trial. Or worse, the trial was unethical because half the women would not receive the “known benefits” of HRT. But many at the meeting¹⁰ commented that the earlier, positive, observational epidemiology studies had looked at groups of women who were fundamentally different from those enrolled in this trial. The women who were taking HRT in the observational studies were generally leaner, less likely to smoke, better educated, more likely to exercise, and more likely to seek medical care than those who did not take HRT. Moreover, when the U.S. Preventive Services Task Force, an independent panel funded by the Agency for Healthcare Research and Quality (AHRQ), conducted a systematic review of HRT studies, applying the rigorous criteria developed for weighing evidence, it found “insufficient evidence” to support long-term use of HRT for chronic disease prevention.¹¹ After the results of the WHI became available, this conclusion was changed to a recommendation *against* use of combined HRT for the prevention of chronic disease.¹²

A careful profiling of subjects in the observational studies that revealed a selection bias, and a rigorous systematic review of the literature that indicated insufficient evidence, thus might have blunted objections to launching a large-scale, definitive, randomized controlled trial (RCT) such as the WHI. A further caveat emerging from discussion of this and other studies is that more attention must be paid to safety in clinical trials, in addition to the usual focus on efficacy. That is, more scrupulous design and evaluation are needed in the smaller Phase 1 and Phase 2 clinical trials that normally precede Phase 3 trials and are conducted to determine dosage, safety, and efficacy.

Many of these same issues arose in “M and M” reviews of other clinical evidence. Based on reports that people who eat foods rich in beta carotene had lower risks of cancer, including lung cancers, and the assumption that the antioxidant effects of the beta carotene or of carotenoids in general were responsible for this effect, at least four RCTs were conducted in the United States and abroad.^{13,14,15,16} These RCTs varied in length,

dosages, and the carotenoids employed. Overall, the trials showed that beta carotene had no benefit in well-nourished populations. A few studies actually showed an increased risk of lung cancer incidence and mortality in smokers exposed to the nutrient. Some differences in health outcomes in these trials could be accounted for by differences in the health and nutritional status of the subjects. For example, poorly nourished Chinese subjects in one trial showed some benefit from beta carotene combined with other vitamins and minerals.¹⁶ Subsequent mechanistic studies, showing a wide array of molecular responses to beta carotene, might account for the adverse findings. Further analysis of beta carotene itself also indicates that, at best, it is a weak antioxidant and in high dosages becomes a pro-oxidant.¹⁷ Some now question whether it ever makes sense to tie broad health benefits (cancer prevention, good for the heart, etc.) to an isolated factor or micronutrient in food. The touted benefits may depend on the complex mix of nutrients in food and the adequacy of the food supply.

The beta carotene trials overturned powerful intuitions about the role of specific nutrients in cancer prevention. Workshop participants also voiced concern about unwarranted assumptions that could lead to delays in conducting a trial. For example, for more than a decade, beta blockers were not prescribed in cases of chronic heart failure. It was assumed that the drug's actions in lowering blood pressure and slowing heart rate would worsen heart failure. In fact, these drugs are beneficial, and RCTs have proven that beta blockers decrease mortality in patients who suffer chronic heart failure.^{18,19,20,21,22} In this case, delay in clinical trials led to delay in implementing a simple intervention now known to be beneficial for a common life-threatening condition.

Biomarkers Defined

The shift in predominant patterns of disease—from acute infections to cancers, heart disease, diabetes, and other chronic degenerative diseases affecting older populations—has expanded the market for new drugs that not only can be promoted to health professionals but also advertised directly to consumers. This shift also reflects a change in FDA rules in 1997. Furthermore, the market for new drugs has created a growth industry in biomarkers. Broadly defined, a biomarker is “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.” The biomarker could be a molecule found in serum or tissue, a gene, or some other physiological indicator, such as a brain image.

Biomarkers as Screening Tests

A biomarker associated with a risk for disease is a candidate for a screening test. For example, the prostate-specific antigen (PSA) can be detected in serum and has been widely hailed as a biomarker for prostate cancer. Tests showing high PSA values in patients result in clinical decisions for biopsy and subsequent medical or surgical treatment. But do patients benefit? The question remains open, but the harms are inescapable, as one major clinical trial showed.

The Prostate Cancer Prevention Trial (PCPT). This placebo-controlled RCT enrolled nearly 19,000 men, aged 55 or older, who had PSA scores of 3 or less.²³ The men in the trial were treated with the anti-testosterone drug finasteride, or a placebo, and followed with periodic PSA tests and digital rectal exams for 7 years. All were offered a biopsy at the trial's end. The prevalence of prostate cancer in the finasteride group was 18.4 percent compared to 24.4 percent in the placebo group ($p < 0.001$). However, more than one-half of the men in the placebo group who were found to have cancer had a normal PSA level and normal digital rectal examination throughout the trial. In a followup study of men in the placebo group, investigators concluded that no cutpoint of PSA is such that higher scores are associated with higher risk for prostate cancer and lower scores with lower risk. Too many false positives and false negatives occur at every score.²⁴

The PSA test provides some important object lessons: it can detect a broad spectrum of prostate cancers. However, prostate cancers are a heterogeneous group—ranging from rapidly progressing and aggressive to slowly progressive or nonprogressive. Treating the slowly progressive or nonprogressive group may lead to the erroneous conclusion that these patients were “cured” by the screening and subsequent treatment. If not screened or treated, these same men might simply have died from other causes—at the same point in time. Early diagnosis based on screening also may lead to the assumption that screening increases survival time, since survival is measured from the time of diagnosis. In reality, it may make no difference in mortality, just a difference in the length of time between the date of diagnosis and death—longer in the case of the screening diagnosis compared to the time when the patient becomes symptomatic (so-called “lead time” bias).

Biomarkers as Surrogate (Intermediate) Endpoints

In addition to their use in screening tests, biomarkers also can serve as “surrogate” or “intermediate” endpoints, instead of the true health or clinical outcome of a trial (ultimately, how the patient feels or functions, or whether he or she survives). A surrogate endpoint is a biomarker thought to be in the pathogenic pathway of disease and, at least theoretically, able to give the same inference for judging interventional efficacy as the health outcome. The virtue of a surrogate endpoint is that it provides a window at an intermediate point t in the trial, short of the true health or clinical outcome, and can serve as a bellwether—indicating whether treatment x is working or not, thus saving both time and money. But that depends on the validity of the surrogate. Ideally, one would like to establish that the relationship between surrogate S and intervention x provides valid information on the association between x and the true outcome. But validity is not easy to establish. Indeed, the selection of biomarkers as surrogate endpoints is often problematic, as illustrated in the following examples.

Diabetes Control and Complications Trial (DCCT). Diabetic retinopathy, with severe visual impairment or blindness, is one of the long-term sequelae of diabetes. The DCCT was a randomized, multicenter trial of 1,441 diabetic patients to determine, among other things, whether intense monitoring of glucose and the use of an insulin pump would reduce the risk of retinopathy.²⁵ The development of microaneurysms was chosen as a surrogate marker, because they are associated with visual loss. Early trends indicated an

increase in microaneurysms and could have led to premature termination of the trial. However, longer term followup showed definite reduction in visual impairment, so the trial was appropriately ended at that point, since it demonstrated a health benefit.

Cardiac Arrhythmia Suppression Trial (CAST). This double-blind, placebo-controlled trial was conducted to determine the benefits of drugs that reliably suppress cardiac arrhythmia in heart patients.²⁶ Because cardiac arrhythmias are strongly associated with risk of sudden death, it was assumed that suppression of arrhythmias would lower the risk of death. However, the trial was terminated early when, among patients taking one or another of the test anti-arrhythmia drugs, sudden deaths occurred at more than triple the rate of those on placebo, and overall mortality was 2.5 times greater. These results occurred despite the fact that the active drugs clearly decreased cardiac arrhythmias.

Similar cautionary tales result from the use of surrogate endpoints in other trials. A trial comparing continuous oxygen therapy with oxygen given only at night to patients who had chronic obstructive pulmonary disease used a number of surrogate markers (e.g., hematocrit, cardiac index, pulmonary vascular resistance) to monitor effects. The trial seemingly indicated no deleterious effects of the nighttime-only regimen, as assessed by the surrogate endpoints. Nevertheless, the patients who received continuous oxygen therapy had lower mortality than the group on the nighttime-only regimen.²⁷ In each case, the surrogate endpoint gave a qualitatively different outcome from the true health outcome.

Short-Term Trials With Long-Term Consequences

Many clinical trials today aim either at preventing rather than curing disease or at forestalling disease progression and late-stage complications in people who have chronic diseases. Because such trials often require very large sample sizes and long followup to achieve definitive health outcomes, they are very likely to use surrogate (intermediate) endpoints to decrease trial size and duration. The results of a clinical trial of limited duration in relatively healthy adults can then translate to a recommendation for lifetime use of a drug (e.g., statins, antihypertensives) by an individual who may be older and less fit than the trial subjects and who may be taking numerous drugs for other chronic illnesses. Concerns about the lifetime use of drugs, especially in already compromised patients, have led many to urge longer trials, or at least longer term followup, empowering the FDA to conduct postmarketing surveillance and to disseminate broadly information on adverse events. This change would require substantial refinement of postmarketing surveillance tools. Currently, these are ill-suited to detect any but very strong “signals” of serious harms with long-term medication use in the general population. Gone are the days when the safety and efficacy of a drug like penicillin could be established rapidly and unequivocally by “the simple ocular test” in which the result stares you in the face! Now, absolute increases in benefit are small, side effects in relatively healthy people must be kept to a minimum, and the “signal-to-noise” ratio of trial results is much lower than when penicillin was discovered.

The Hippocratic Injunction Still Applies: First, Do No Harm

Not surprisingly, one of the strongest messages emerging from the NIH meeting was that surrogate endpoints must be used with caution. All too often it is assumed that these endpoints are *causally* associated with the clinical or health outcome and can be substituted for the health outcome in drawing conclusions. In fact, a surrogate's association may be coincidental; it may be associated with only one of several pathways (and not the most important) in the pathogenesis of disease; or it may be unrelated to or unaffected by the intervention, which may act on another pathway. Indeed, the surrogate may have a mode of action little understood in relation to the disease process.²⁸ Similar caveats apply to biomarkers selected as screening agents. Too often, it is assumed that biomarkers are nearly pathognomonic of the disease in question and that their use can advance the date of diagnosis, leading to a virtual certainty of benefit. However, it is important to keep in mind that “earlier detection” is not synonymous with “early detection” or “early enough detection.”

Validating Biomarkers and Surrogates

Recognizing the potential value of biomarkers and surrogate endpoints in clinical trials, how best can they be validated, especially when it is not feasible to extend a trial long enough to reach a true clinical endpoint? Among methods proposed at the meeting was the use of a set of formulas involving the statistical concept of hazard ratio. That is, for health risks that have a hazard ratio greater than 1, the chances of getting the health risk *increase* with the treatment. If the ratio is less than 1, the chances of getting the health risk *decrease* with treatment. This framework could be used to establish, at the strongest, a causal link between a surrogate and the true clinical or health endpoint, or—less stringently—a strong association.²⁹ Also strengthening the case for validity would be corroborative findings from meta-analyses of multiple small trials of a surrogate in relation to a given therapy. Indeed, evidence from such trials was deemed “fundamental to the development and screening of preventive interventions” for subsequent testing in a definitive clinical trial. (Meta-analyses are a means of obtaining greater statistical power and precision for data analysis by compiling results from smaller clinical trials or studies as long as the data meet the analysts' selection criteria and the studies allow for comparability.) Other tests of validity for biomarkers invoke statistical measurements indicating that the marker demonstrates high sensitivity and specificity, or has high predictive value, in an independent test sample.³⁰

To validate surrogate endpoints, a “validity” trial in which both surrogate and true endpoints are observed should be conducted. The trial should be one in which it can be concluded that the effects of the intervention were the same, whether based only on the surrogate or only on the true endpoint. It is important that the surrogate be shown to be truly on the causal pathway to the clinical outcome and that, as the degree of exposure to the surrogate increases, so does the likelihood of the clinical outcome. Indeed, the marker mediates the association between exposure and outcome.

Other methods for the validation of surrogates include an estimation framework using meta-analyses, a trial applying the surrogate to predict the effect on the true endpoint, and two meta-analytic methods for combining data from previous trials. Unfortunately, to date, these various frameworks rarely, if ever, actually are employed to validate most surrogate endpoints, including surrogates used in cancer studies.

Can New Technology Make Trials Better?

Learning from past clinical trials is critical today, not only because of the explosion in medical advertising and escalating pressures on journals to publish “exciting” findings, but because of advances in science. The 21st century is seeing the rapid expansion of research on risk factors for disease and a growing public interest in “wellness” and disease prevention. The tools for this research are the new technologies and databases of the “-omics” revolution, including genomics, proteomics, and metabolomics. Researchers conducting observational and epidemiologic studies can collect tissue samples to discover which genes or proteins may be associated with the risk for cancer and other devastating diseases—or conversely may confer a lowered risk. But concerns have surfaced about the credibility of positive findings in such needle-in-a-haystack searches, leading some statisticians to propose rethinking such venerable measures of significance as a p value <0.05 and to consider refinements in data analysis to guard against false positive results.³¹ The words of caution to apply here are that when searching through a haystack for needles, most sharp objects are pieces of straw.

The use of microarray technologies to determine patterns of genes activated in cells affected by disease compared with those in normal cells is a case in point. This is an exciting tool, but one to be used with care, given the very human tendency to find patterns where none exist and the potential that multivariable analysis can lead to specious results.^{32,33} At the very least, independent replications are essential. As participants at the NIH meeting were admonished, the tools and technologies of scientific investigation grow ever more sophisticated, but the rules of evidence have not changed.

Go or No Go?

Uncovering confounders in past clinical studies, picking good biomarkers/surrogates, and clarifying the difference between causal and noncausal associations can help “get trials right” in the future, but that still leaves open the question of whether or not to do a trial in the first place. At stake are issues that go to the heart of the medical enterprise. The goal of a clinical trial is to benefit public health, usually by testing an intervention to reduce the morbidity or mortality of a disease.

Therapeutic Milestones

Toward that end, medical history provides successive milestones in discovery by pioneers who made the right connection between cause and effect: Lind on scurvy, Semmelweis on puerperal fever, Snow on cholera, Pasteur on the germ theory. All these innovators advanced understanding of the causes of diseases and of how to diagnose, treat, and prevent them. However, not until the 20th century did medicine codify a need for rational

therapeutics, and successive laws were passed to assure the purity of foods (1906) and the safety (1937) and efficacy (1962) of drugs and devices. Meanwhile, advances in probability theory and statistics, especially the work of R. A. Fisher³⁴ and others, created the scientific underpinning for the design and analysis of controlled clinical experiments—the most efficient means to detect cause-and-effect relationships. Today’s large, double-blind, placebo-controlled RCT is considered the “gold standard” for establishing the balance between benefits and harms of a therapeutic intervention. Each qualifier—the number of subjects, their randomization, the blinding of investigators and subjects, and the use of placebo controls—contributes to a protocol designed to generate a set of adequate data to accept or reject the trial’s hypothesis, while eliminating bias and other confounding factors, thus enabling confidence in the translation of findings into clinical or public health recommendations.

Often, however, the decision to mount a clinical trial is made late in the overall study. Priority setting is a necessity, given the expense, duration, and potential for harm as well as benefit of a Phase 3 trial; the myriad new drugs, biologics, and devices that emerge from the Nation’s laboratories every year; and the limited amount of funds available. Although it is all well and good to state that a trial will address the magnitude of a health problem or contribute to scientific understanding, other priorities clearly enter into decisions as to whether to launch a definitive clinical trial. Among the priorities are those listed in table 1.

Table 1
Factors Affecting the Decision To Conduct a Clinical Trial

- Strength of existing evidence: promising, but not conclusive data from preclinical and observational studies/small trials
- Potential impact on public health and/or medical science
- Therapeutic equipoise: impossible to decide which of two therapies is better without a head-to-head test
- Portfolio balance: a research organization may adjust its priorities to give more (or less) attention to specific areas to achieve optimal use of its resources
- A potential for runaway practice, e.g., a new diagnostic test that is widely marketed but has not been adequately evaluated
- Social/political context/pressures—from legislators, community leaders, patient advocacy groups

What Is the Evidence? What Is the Action?

Other criteria aside, the strength of existing evidence is the most obvious and logical basis for the decision to move to a clinical trial. Indeed, in the last decades of the 20th century, the medical community as a whole expressed a commitment to “evidence-based” practice.³⁵ But what constitutes *good* evidence of causation/association or evidence *strong enough* to justify a trial or clinical/public health action is often arguable. It is generally agreed that the well-designed large RCT, double-blind and placebo-controlled, is at the top of a hierarchical ladder of evidence. It is followed, in order of diminishing strength, by the evidence from smaller RCTs, uncontrolled trials, observational studies, case studies, and—at the bottom—logical constructs (opinion) and anecdotes.³⁶ Whatever data exist at any level in the hierarchy need to be weighed for merit or bias in the process of decisionmaking.

However, this analysis of the literature often involves other considerations, such as the cost and cost-effectiveness of the intervention, third-party coverage, the level of risk one is willing to tolerate for the potential benefit, whether side effects are minor or major, political influences, the potential profit from a new drug, pressure from advocacy groups, and so on. As a result, even in the presence of large data sets, ultimate decisionmaking and policy making may not always be completely logical, consistent, and transparent, and certainly will reflect subjective values and judgments.

These worrisome issues triggered considerable discussion at the NIH meeting, beginning with examples of formal approaches for weighing evidence leading to the conclusions about association or causation. Koch’s postulates were among the earliest examples of formal rules to apply to determine whether a candidate microorganism causes an infectious disease. These rules were used most recently in determining the viral cause of SARS.³⁷ In the case of chronic diseases, more complex guidelines for causation have been put forward, based on the appreciation that chronic diseases are caused by multiple and often interacting variables. In the mid-1960s, Sir Austin Bradford Hill proposed a list of nine criteria to be applied to observational studies to cite causation: the strength of the association (the larger the better), consistency (do independent studies agree?), specificity (this implies a one-to-one cause–effect connection and is not a necessity—or even frequently the case), temporality (the cause precedes the effect), biological gradient (a dose–response relationship), plausibility (one could conjecture a mechanism of action), coherence (the association makes sense from what is known about the natural history), experiment (can one be designed to test the proposed cause?), and analogy (is the cause–effect relationship comparable to other known associations?).³⁸ Variations and subsets of the Hill criteria are commonly used today to judge associations/causality in epidemiologic studies and in making decisions to undertake clinical trials.

The focus on criteria for causality occasioned a discussion of what surrogate endpoints might be “good for.” Indeed, they might contribute to biological plausibility. For example, the presence of *H. pylori* as a surrogate (intermediate) endpoint is now well established as essential for the development of stomach ulcers. Intermediate endpoints can also enhance a causal connection by showing a dose–response relationship

(increasing strength of exposure → increased disease association) as well as clarifying causal pathways. In studies of gene–environment interactions, undertaken to predict risk for disease, a gene marker can be used, in combination with the environmental factors, to sharpen relative risk for disease. In one study cited, selected allelic variants of genes encoding for metabolizing enzymes were used, along with measures of daily intake of red meat, as predictors of risk for colorectal cancer. The study demonstrated that, in combination, these intermediate indicators were associated with a fivefold increase in relative risk among 212 individuals in the Physicians’ Health Study who developed colorectal cancer.³⁹ Nevertheless, being on the causal pathway does not guarantee the validity of the surrogate endpoint in a clinical trial as an accurate predictor of or substitute for the health outcome.

Meta-analyses and Systematic Reviews

The two formal tools most commonly used today to obtain an accurate overview of the literature are meta-analyses (described earlier) and systematic reviews. Systematic reviews have been pioneered by such organizations as the Cochrane Collaboration,⁴⁰ the AHRQ,⁴¹ and the AHRQ’s U.S. Preventive Services Task Force.⁴² For each systematic review, the reviewers develop a set of criteria and grade the strength of each study selected for review. The individual grades are then looked at in the aggregate to determine the overall level of evidence: e.g., strong, weak, insufficient.

Even with the use of these more objective approaches, however, there is a danger that study biases can tilt the outcome of a systematic review. One would hope that finding consistent results of multiple meta-analyses of smaller nonrandomized studies would be confirmed in large clinical trials. However, data from a dozen large (more than 1,000 subjects) RCTs showed that 35 percent of the time, the outcomes were not predicted accurately by meta-analyses published previously on the same topics.⁴³ Even the outcomes of megatrials themselves on the same topic are not always consistent. In one study of pairs of large trials on the same topic, results of 79 of the 289 pairs (27 percent) differed in statistically significant ways.⁴⁴

Given these discrepancies, one is tempted to conclude with the cliché, “more research is needed.” In effect, each mode of weighing the clinical evidence has its strengths and weaknesses and can contribute its weight to decisionmaking. Large trials offer strengths through sheer numbers and a single protocol. On the other hand, they may suffer from insufficient generalizability. Meta-analyses offer the possibility of increased statistical precision. On the other hand, they do not necessarily ensure greater validity but may just increase the precision of a biased estimate of benefit or harm. Meta-analyses are, by nature, observational studies subject to systematic biases or confounding.

Communicating the “Truth”

Once a clinical trial has been done and the results are in, questions arise regarding when and how the results are to be made public. It is well known that some sponsors are reluctant to publish negative or adverse findings—as are the peer-reviewed journals in

some cases—although signs indicate that change is on the way. Recently, the major medical journals have agreed to publish articles on clinical trials only if they have been previously registered in an accessible database.⁴⁵ Another issue that may delay release of information is the need to inform professional societies of findings that may change medical practice. Still another influence on trial sponsors may be the anticipated effects on the stock market.

In any case, the results of clinical trials are of little benefit to the public, patients, practitioners, or policy makers unless they are clearly reported in a timely fashion. But what is said sometimes reveals a spin on data to persuade or to favor a particular point of view. Results of a new therapy may be stated in terms of a *relative* percent improvement, but a percent of what? The base rate—the number of clinical events or the event rate—is often not stated. This means of reporting data often exaggerates apparent benefits compared to reporting *absolute* rates with, versus without, an intervention or exposure. Results also may be couched in statistical calculations, like odds ratios, that can generate impressive numbers. Many people mistake odds ratios for relative risks. If the event rate is rare, this is not a problem, but as the outcome becomes more common, the odds ratio gets bigger and bigger compared to the relative risk. Over the years, books like *How to Lie With Statistics*⁴⁶ and *News and Numbers*⁴⁷ have been instructive in explaining the ways data can be manipulated. If anything, the need for such primers has grown with direct advertising of prescription drugs to consumers. Several speakers provided telling examples of data “framing”—not technically incorrect, but misleading—that have been used in medical advertising and even in articles in peer-reviewed journals.

For example, an ad for an anti-osteoporosis drug claimed that individuals taking the drug in a year’s trial experienced a 68 percent reduction in clinical vertebral fractures over a comparable group on placebo. However, a look at the actual figures (see sidebar) showed that there were 5 fewer fractures per 1,000 women among the drug users—leaving a far different impression of the magnitude of benefit. Medical ads also typically downplay side effects by placing them in the small print, usually with an aside to “check with your doctor.” In some cases, relative rates are given for benefits, but absolute rates are given for side effects.

... a 68% reduction!

The ad for an anti-osteoporosis drug claimed that individuals taking the drug in a year's trial experienced a 68% reduction in clinical vertebral fractures compared with a similar group taking placebo.

What were the actual figures?

The rate of fractures among placebo users was .738%.

The rate of fractures among the drug users was .238%.

Thus, the absolute risk reduction was

$$.738\% - .238\% = .5\%$$

which translates to 5 fewer fractures per 1,000 women—hardly headline news.

But by presenting the data in terms of relative risk (RR) reduction,

$$\text{RR reduction} = 1 - .238\% / .738\% = .678 \text{ (~68\%)}$$

the impressive 68% figure was advertised. Computations of relative risk reduction will always appear impressively large when actual event rates are low.

Similarly, a peer-reviewed paper in a leading medical journal, reporting on a population-based case-control study of breast cancer, stated that women who used aspirin over a 5-year period had a 20 percent reduction in breast cancer—a figure all but guaranteeing widespread coverage in the media.⁴⁸ Here, the 20 percent represents the relative risk, derived by dividing the 1.6 percent risk of breast cancer reported for women who took aspirin by the 2.0 percent risk of breast cancer in women not taking aspirin and subtracting the fraction from 1: $1 - 1.6\% / 2.0\% = 1 - .8 = .2 = 20\%$. Furthermore, the paper did not report harms associated with chronic aspirin use—dangers of hemorrhagic stroke or intestinal bleeding. (A more recent large, randomized, placebo-controlled trial has shown no benefit of aspirin on incidence of breast cancer.⁴⁹)

Some sleuthing may be needed to find the critical event rates and risks associated with data that have been framed to imply dramatic effects. Take for example a study that reported that women and

Blacks who reported chest pain were 40 percent less likely to be referred for cardiac catheterization than Whites or men with the same symptoms. The culprit here was the use of odds *ratios*, a poor approximation of relative rates in this case, and misleading when absolute rates of referral were examined. (See the sidebar for the actual figures.)

Other ways of obscuring or presenting misleading information involve the use of scores.

The odds of referral for cardiac catheterization:

The actual rates for referral were:

85% for Blacks

91% for Whites

Thus the odds (the ratio of occurrence to nonoccurrence) for referral were:

$$\frac{85}{15} = 5.5 \text{ for Blacks, and } \frac{91}{9} = 9.6 \text{ for Whites,}$$

making an odds *ratio* of

$$\frac{5.5}{9.6} = 0.6$$

This was interpreted to mean that Blacks (and women) were *40 percent less likely to be referred for catheterization than Whites*. If relative risk had been calculated, the result would have shown that Blacks or women are referred 7 percent less often than

Whites or men ($1 - \frac{85\%}{91\%} = 1 - .93 = .07 = 7\%$).

Typically, these are ways of summing up responses to a set of related questions about the effects of an intervention. One study cited scores used by physicians to rate the effects of a drug on the functioning of an individual with Alzheimer’s disease. The reports did not clarify what the scales measured, the range from low to high, or how significant a difference in a few points up or down the scale might be in the functional status of the patient.

The “take-home messages” from these analyses are to search for the meaning of the data behind the headlines—seek out actual event rates, for both benefits and harms, translate odds ratios into relative risks, get the meaning of scores—and to have

investigators use more useful statistics in their reports in peer-reviewed literature.⁵⁰

The Media and Advocacy Groups

That message is taken very seriously by two other key players in communicating the results of clinical trials: reporters and patient-advocacy groups. The best of them do much more than act as passive conveyors for news releases of scientific findings. They check the sources and talk to the principal investigators—and often to their peers and competitors. In other words, they serve as “honest brokers” of the evidence and provide context for new findings. This was the thrust of remarks made at the meeting by a health and science journalist for network television. He described the deluge of e-mails, phone calls, and even videotapes he gets every week promoting a health story, and, in turn, the pressures reporters face from their editors to report certain stories “no matter what.” The reputation of NIH remains outstanding, he said, and attention is paid to any NIH news release. However, he warned that NIH needs to do all it can to assure the public of its integrity in interpreting and reporting data lest that faith be shaken.

Patient advocacy groups play a different role than journalists, because they have a vested interest in promoting research on specific diseases/disorders. Advocacy leaders often maintain direct contact with agencies sponsoring trials, conduct oversight, and provide

feedback. Sometimes a group has been responsible for setting a clinical trial in motion in the first place and then helped recruit volunteers for participation. Occasionally, the opposite occurs: the group will oppose a trial or be split down the middle. Noteworthy among such groups has been the AIDS activist organization, Act Up. A representative of that organization described actions the group has undertaken over the years. Early on, he recounted, the group had seized on results of early trials of azidothymidine (AZT) that indicated the dosages prescribed were too toxic. The group contacted NIH and FDA officials and initiated a congressional hearing that resulted in swift action to lower the recommended dosage. He also had a cautionary tale about surrogate markers. To speed development of AIDS drugs in the early 1990s, some advocates urged use of an increase in CD4 T-cell counts as a surrogate for efficacy in clinical trials. For dideoxyinosine (DDI), this approach was a mistake. He also emphasized the importance of keeping channels of communication open, especially when agencies conduct clinical trials abroad. He cited an AIDS trial in Cambodia that was shut down because of failure to involve the local community at all stages of planning the trial. One of his most important points related to ethical considerations in conducting a trial in which volunteers accept the possibility of getting the less effective of two possible approaches. In an AIDS trial using clinical endpoints to determine the efficacy of combinations of three versus two AIDS drugs, the higher mortality rate in the two-drug group rapidly led to the establishment of the three-drug regimen as the global standard for treating HIV/AIDS.

Prospects for the Future

The AIDS trials, which led to today's life-saving drug regimens for a disease all but universally fatal 25 years ago, are just one example of the value to health and well-being of well-designed, well-conducted clinical trials. These trials underscore why a meeting to learn "Why Do We Sometimes Get It Wrong?" was so important. Despite the litany of shortcomings in observational studies that led to failed trials, an overall air of optimism predominated. That is, taming the new technologies and applying strategies for rigorous statistical and study design, to provide greater assurance against the probability that a positive finding may be false, may help resolve some of the uncertainties that bedevil clinical trials today. Greater knowledge of pharmacogenetics and metabolomics, as well as better understanding of how genes interact with each other and with environmental variables, may lead to new conceptual designs for clinical trials in the future. In time, there may be fewer instances of anomalous results, uncertain dosages, and adverse events that result in terminating a trial. Medicine may see an acceleration of the movement toward prevention of disease and maintenance of health and well-being. All this will be welcome news to the public, advocacy groups, the media, the policy makers—and scientists themselves, who are among the first to admit, as Aristotle said in *Metaphysics II*, "The investigation of truth is in one way hard, in another easy."

Summary and Recommendations

The January 11–12, 2005, meeting at NIH, *Moving From Observational Studies to Clinical Trials: Why Do We Sometimes Get It Wrong?* convened by NIH Director Elias A. Zerhouni, M.D., challenged experts to pinpoint sources of error in recent observational

studies and clinical trials and to indicate how to avoid such errors in the future. In addition, the group was asked to consider ways to maximize the quality of evidence, within budgetary restraints, and to identify methodologies or frameworks currently available or needed to judge evidence in today's research environment. An underlying theme was to explore research directions to validate methods that distinguish causal and noncausal associations.

The overarching message of the meeting was that investigators must do all they can to avoid the easy assumptions, mental shortcuts, intuitive beliefs, and conventional but unfounded "wisdom" that have led research astray—as illustrated by the unexpected serious adverse outcomes in several recent large clinical trials. The rules of evidence have not changed, and these rules demand that attention be paid to sources of error at every step in proceeding from clinical observations to Phase 3 clinical trials. Toward that end, the authors developed the following set of recommendations, based on discussions at the conference.

Recommendation 1. Eliminate bias and confounders to the extent possible. Have a healthy respect for the power of bias and confounders, and search diligently for them.

Selection bias, nonblinded data analysis, inadequate sample size, etc.—all the standard dos and don'ts taught in basic statistics—apply in weighing the evidence at every step in the hierarchy of evidence, from anecdotes and case studies to nonblinded observational studies to small clinical trials. Yet these factors have been overlooked and found to be hidden but powerful sources of error in recent clinical studies when put to the definitive test of an RCT.

Recommendation 2. Establish formal means of setting priorities for large clinical trials. Two categories were put forward as being of high priority in deciding to conduct large clinical trials: the potential benefit to public health, as defined by the magnitude and gravity of the health problem; and the contribution the clinical trial can make to fundamental scientific knowledge and understanding. Indeed, on occasion, both criteria may pertain. A trial providing definitive results for an intervention in a relatively rare disease may inform a wide array of problems or shed light on new mechanisms of pathogenesis or therapy. A third criterion, which might override the first two, is the case of a runaway practice, e.g., a screening diagnostic method advocated for widespread use in the absence of any validation of safety or effectiveness. It is important that decisions to conduct trials be logical, consistent, and transparent.

Recommendation 3. Give equal weight to determining safety as well as efficacy in clinical trials.

Some problems associated with adverse outcomes of large clinical trials might have been averted had greater attention been paid to issues of safety in observational studies and in Phase 1 and Phase 2 clinical trials. These considerations take on added significance in light of the fact that an intervention may be shown to be safe and efficacious in a Phase 3 trial of relatively healthy volunteers. This result may lead to prescribing a drug for

lifelong use in older individuals whose health may be compromised and who already may be taking drugs for other chronic illnesses. To provide greater assurance against adverse events, the group proposed that some trials be extended for much longer periods. Importantly, the group recommended that, once a drug has been approved by the FDA and been made available to the public at large, the FDA should be empowered with the necessary technology and resources for conducting long-term surveillance, as well as to enhance sensitivity and specificity of the surveillance tools.

Recommendation 4. Use meta-analyses and systematic reviews to enhance means of weighing evidence, but beware of potential systematic biases.

Meta-analyses and systematic reviews have become valuable and accepted means of obtaining greater weight and precision in judging clinical evidence. Nevertheless, these methods reflect the selection and evidential criteria of the analysts and reviewers, and the analyses essentially turn the trials into observational studies. Multiple meta-analyses and multiple reviews can mitigate the potential bias of singleton studies. Even in those cases, however, results (reported for meta-analyses) may not be consistent when compared with results of large “definitive” clinical trials on the same topic.

Recommendation 5. Validate biomarkers and surrogate endpoints before basing policy guidelines for public health on them.

Biomarkers and surrogate endpoints have become essential tools of the trade in setting priorities for or conducting large-scale clinical trials. Researchers are motivated to use them in part because they are means of limiting the duration and expense of Phase 3 trials. However, the selection of biomarkers and surrogate endpoints may reflect anecdotal/intuitive judgments about cause–effect or other strong associations. Methods of validating markers and surrogates are critical and specifically include the testing of the candidate marker in a trial that extends to the actual health or clinical outcome. Several frameworks can be used to judge causality and association.

Recommendation 6. Use the technologies of the “-omics” revolution and systems biology approaches, but use them with more caution than has been used on occasion.

The potential of exploiting genomics, proteomics, pharmacogenetics, metabolomics, and other “-omics” data, in combination with a systems biology approach that looks at the various cell/tissue/organ/environmental interactions, may truly revolutionize medicine in the 21st century and lead to new conceptual designs of clinical trials or medical interventions. For example, although such studies are still in their infancy, researchers are already using microarrays to search for specific genes or gene clusters activated in cancerous or other diseased cells. The pitfalls of such searches were noted at the meeting—specifically, the human tendency to find patterns where none exist. Proof of principle is not yet in hand for some of these promising technologies. To guard against false positives, some change in thinking may be needed, including rethinking the use of p values and other refinements in statistical analysis.

Recommendation 7. Communicate results of clinical trials in an accurate and timely manner. Include basic information about event rates; use absolute rates when possible, not simply relative rates; and use other means as necessary to make clear—to the intelligent layperson—what the results of the trial mean.

Both the public and health professionals all too often suffer from the provision of imperfect information imperfectly reported, either by failure to report negative results of trials or by manipulation of the data to suggest impressive benefits and insignificant harms. Major medical journals have taken steps to guard against inadequate reporting and have established new rules. One is to stipulate that they will publish only results of clinical trials that have been previously put on record in a publicly available registry. However, it is acknowledged that this will not eliminate the problems of medical advertising directly to consumers. These ads often accentuate the positive and all but eliminate the negative.

Recommendation 8. Establish open and two-way communications with communities, consumers, and patient-advocacy groups in the course of development, implementation, and reporting of clinical trials.

As part of a move toward greater transparency, sponsors of clinical trials can benefit by informing concerned individuals and groups of their plans to conduct large clinical trials and by showing their willingness to cooperate/collaborate with such groups. This action is especially important when U.S. sponsors conduct trials in other countries.

Recommendation 9. Establish criteria to inform public policy and decisionmaking. Given the potential life-saving or life-enhancing benefits of a clinical trial, what are the formal steps needed to move health information or outcomes information into public health policy making and decisionmaking? Again, transparency and consistency in the process are key elements. In Government-sponsored trials, the agency sponsoring the trial bears the responsibility for the timely (1) publication of findings, (2) communication to the public and health professionals, and (3) provision of information to the Secretary of the Department of Health and Human Services and to the heads of other public health agencies as appropriate, including the NIH, the FDA, the Centers for Disease Control and Prevention, the AHRQ, the Substance Abuse and Mental Health Services Administration, and the Centers for Medicaid and Medicare Services. Of course, final responsibility for changes in public health policy rests with elected representatives of the public.

Sources

¹NHLBI stops trial of estrogen plus progestin due to increased breast cancer risk, lack of overall benefit [NIH news release]. 2002 Jul 9. Available from: www.nhlbi.nih.gov/new/press/02-07-09.htm

²Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, et al. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in

healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA 2002 Jul 17;288(3):321-33.

³Merck announces voluntary worldwide withdrawal of Vioxx [news release]. 2004 Sept 30. Available from:

www.vioxx.com/rofecoxib/vioxx/consumer/press_release_09302004.jsp

⁴Eisenberg RS. Learning the value of drugs—is rofecoxib a regulatory success story? N Engl J Med 2005 Mar 31;352(13):1285-7.

⁵NIH halts use of COX-2 inhibitor in large cancer prevention trial. [NIH news release, 2004 Dec 17]. Available from: www.nih.gov/news/pr/dec2004/od-17.htm

⁶Kaufman, M. FDA panel opens door for return of Vioxx. Washington Post 2005 Feb19; Sect. A:1.

⁷COX-2 selective (includes Bextra, Celebrex, and Vioxx) and non-selective non-steroidal anti-inflammatory drugs (NSAIDs) [report from the FDA Center for Drug Evaluation and Research]. Available from: www.fda.gov/cder/drug/infopage/COX2/default.htm

⁸Bodley H. Medical examiner: ephedra a factor in Bechler death. USA Today 2003 Mar 13. Available from: www.usatoday.com/sports/baseball/al/orioles/2003-03-13-bechler-exam_x.htm

⁹Suicidality in children and adolescents being treated with antidepressant medications [FDA public health advisory]. 2004 Oct 15. Available from: <http://www.fda.gov/cder/drug/antidepressants/SSRIPHA200410.htm>

¹⁰Barrett-Connor E. Commentary: observation versus intervention—what's different? Int J Epidemiol. 2004; 33(3):457-9. Epub 2004 May 27.

¹¹U.S. Preventive Services Task Force. Chemoprevention for hormone replacement therapy [summary of recommendations.] 2002. Available from: www.ahrq.gov/clinic/uspstf/uspsmho.htm

¹²What's new from the USPSTF. Postmenopausal hormone replacement therapy for primary prevention of chronic conditions [fact sheet]. 2002 Oct. Available from: www.ahrq.gov/clinic/3rhduspstf/hrt/rtwh.htm

¹³Hennekens CH, Buring JE, Manson JE, Stampfer M, Rosner B, Cook NR, Rosner B, Belanger C, LaMotte F, Gaziano JM, et al. Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. N Engl J Med. 1996 May 2;334(18):1145-9.

¹⁴The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. N Engl J Med 1994 Apr 14;330:1029-35.

¹⁵Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, Keough JP, Meyskens FL Jr, Valanis B, Williams JH Jr, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *N Engl J Med* 1996 May 2;334(18):1150-5.

¹⁶Blot WJ, Li JY, Taylor PR, Guo W, Dawsey S, Wang GQ, Yang CS, Zheng SF, Gail GM, Li GY, et al. Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population. *J Natl Cancer Inst* 1993 Sep 15;85(18):1483-92.

¹⁷Burton GW, Ingold KU. Beta-carotene: an unusual type of lipid antioxidant. *Science* 1984 May 11;224:569-73.

¹⁸Packer M, Bristow MR, Cohn JN, Colluci WS, Fowler MB, Gilbert EM, Schusterman NH. The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. U.S. Carvedilol Heart Failure Study Group. *N Engl J Med* 1996 May 23;334(21):1349-55.

¹⁹Randomised, placebo-controlled trial of carvedilol in patients with congestive heart failure due to ischaemic heart disease. Australia/New Zealand Heart Failure Research Collaborative Group. *Lancet* 1997 Feb 8;349(9049):375-80.

²⁰The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II); a randomized trial. *Lancet* 1999 Jan 2;353(9146):9-13.

²¹Hjalmarson A, Goldstein S, Fagerberg B, Wedel H, Waagstein F, Kjeksus J, Wikstrand J, El Allaf D, Vitarec J, Aldershvile J, et al. Effects of controlled-release metoprolol on total mortality, hospitalizations, and well-being in patients with heart failure: the Metoprolol CR/XL Randomized Intervention Trial in congestive heart failure (MERIT-HF). MERIT-HF Study Group. *JAMA* 2000 Mar 8;283 (10):1295-302.

²²A randomized trial of beta-blockade in heart failure. The Cardiac Insufficiency Bisoprolol Study (CIBIS). CIBIS investigators and committees. *Circulation* 1994 Oct;90(4):1765-73.

²³Thompson IM, Goodman PJ, Tangen CM, Lucia MS, Miller GJ, Ford LG, Lieber MM, Cespedes RD, Atkins JN, Lippman SM, et al. The influence of finasteride on the development of prostate cancer. *N Engl J Med* 2003 Jul 17;349(3):215-24. Epub 2003 Jun 24.

²⁴Thompson IM, Ankerst DP, Chi C, Lucia MS, Goodman PJ, Crowley JJ, Parnes HL, Coltman CA Jr. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower. *JAMA* 2005 Jul 6;294(1):66-70.

²⁵The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. *N Engl J Med* 1993 Sep 30;329(14):977-86.

-
- ²⁶Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med* 1989 Aug 10;321(6):386-8.
- ²⁷Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive lung disease: a clinical trial. Nocturnal Oxygen Therapy Trial Group. *Ann Intern Med* 1980 Sep;93(3):391-8.
- ²⁸Fleming, TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* 1996 Oct 1;125(7):605-13.
- ²⁹Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989 Apr;8(4):431-40.
- ³⁰Baker, SG, Kramer BS, Prorok PC. Development tracks for cancer prevention markers. *Dis Markers* 2004;20(2):97-102.
- ³¹Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004 Mar 17;96(6):434-42.
- ³²Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005 Feb;5(2):142-9.
- ³³Ransohoff DF. Lesson from controversy: ovarian cancer screening and serum proteomics. Comment on *J Natl Cancer Inst* 2005 Feb 16;97(4):307-9 and *J Natl Cancer Inst* 2005 Feb 16;97(4):310-4.
- ³⁴Salzburg D. *The lady tasting tea*. New York: W. H. Freeman and Co.; 2001.
- ³⁵Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't [editorial]. *BMJ* 1996 Jan 13;312:71-2.
- ³⁶Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study [editorial]. *BMJ* 2000 29 July; 321:255-6.
- ³⁷Kuiken T, Fouchier RA, Schutteen M, Rimmelzwaan GF, van Amerongen G, van Riel D, Laman JD, de Jong T, van Doornum G, Lim W, et al. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet* 2003 Jul 26;362(9380):263-70.
- ³⁸Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965 May;58:295-300. Available from: <http://www.edwardtufte.com/tufte/hill>
- ³⁹Chen J, Stampfer MJ, Hough HL, Garcia-Closas M, Willett WC, Hennekens CH, Kelsey KT, Hunter DJ. A prospective study of N-acetyltransferase genotype, red meat intake, and risk of colorectal cancer. *Cancer Res* 1998 Aug 1;58(15):3307-11.

⁴⁰See the Web site: www.cochrane.org

⁴¹See the Web site: www.ahrq.gov

⁴²See the Web site: www.ahrq.gov/clinic/uspstfix.htm

⁴³LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* 1997 Aug 21;337(8):536-42.

⁴⁴Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. *J Clin Epidemiol* 2000 Dec;53(12):1103-9.

⁴⁵Major medical journals have agreed to a requirement that clinical trials be registered in a public registry. For example, the *Annals of Internal Medicine* requirements stipulate that as of 13 September 2005 the journal will consider only registered clinical trials for publication. Before that date, registration is highly recommended but not mandatory. Trials that begin enrollment on or after July 2005 must be registered at or before the onset of patient enrollment. The journal will accept retrospective registration (registration after enrollment begins before 1 July 2005).

⁴⁶Huff D. *How to lie with statistics*. New York: W. W. Norton; 1954.

⁴⁷Cohn V, Cope L. *News and numbers*. 2nd ed. Ames (IA): Iowa State University Press; 2001.

⁴⁸Woloshin S. Comment. *JAMA* 2004 May 26;291(20):2488-9.

⁴⁹Cook NR, Lee IM, Gaziano JM, Gordon D, Ridker PM, Manson JE, Hennekens CH, Buring JE. Low-dose aspirin in the primary prevention of cancer: the Women's Health Study: a randomized controlled trial. *JAMA* 2005 Jul 6; 294 (1):47-55.

⁵⁰Schwartz L, Woloshin S. The media matter: a call for straightforward medical reporting [editorial]. *Ann Intern Med*. 2004 Feb 3;140(3):226-8.