Heuristic Thinking and Inference From Observational Epidemiology

Timothy L. Lash

Abstract: Epidemiologic research is an exercise in measurement. Observational epidemiologic results usually include a point estimate, a measure of random error such as a frequentist confidence interval, and a qualitative discussion of study limitations. Without randomization of study subjects to exposure groups, inference from study results requires an educated guess about the strength of the systematic errors compared with the strength of the exposure effects. Although quantitative methods to make these educated guesses exist, the conventional approach is qualitative, which reduces the educated guessing to a problem of reasoning under uncertainty. In circumstances such as these, humans predictably reason poorly. Heuristics and resulting biases that simplify the judgmental tasks tend to underestimate the systematic error, underestimate the uncertainty, and focus the inference on the study's specific evidence while excluding countervailing external information. Common warnings to interpret results with trepidation are an ineffective solution. The methods that quantify systematic error and uncertainty challenge the analyst to specify the alternative explanations for associations that are otherwise too readily judged causal.

(*Epidemiology* 2007;18: 67–72)

Epidemiologic research is an exercise in measurement. Its objective is to obtain a valid and precise estimate of either the occurrence of disease in a population or the effect of an exposure on the occurrence of disease. Conventionally, epidemiologists present their measurements in 3 parts: a point estimate (eg, a risk ratio), a frequentist statistical assessment of the uncertainty (eg, a confidence interval, but also sometimes a *P* value), and a qualitative description of the threats to the study's validity.

Without randomization of study subjects to exposure groups, point estimates, confidence intervals, and *P* values lack their correct frequentist interpretations.[1] Randomization and a hypothesis about the expected allocation of outcomes—such as the null hypothesis—allow one to assign probabilities to the potential outcomes. One can compare the observed association with this probability distribution to estimate the probability of the observed association, or associations more extreme, under the initial hypothesis. This comparison provides an important aid to causal inference,[1] because it provides a probability that variation in the outcome distribution is attributable to chance as opposed to the effects of exposure. The comparison is therefore at the root of frequentist statistical methods and inferences from them. When the exposure is not assigned by randomization, as is the case for observational epidemiologic research, the comparison provides a probability that the observed association is attributable to chance as opposed to the combined effects of exposure and systematic errors. Given an association with a low probability, a causal inference would then require an educated guess about the strength of the systematic errors compared with the strength of the exposure effects.

These educated guesses can be accomplished quantitatively by likelihood methods,[2] Bayesian methods,[3] regression calibration,[4] missing data methods,[5,6] or Monte Carlo simulation,[7–9] (see Greenland[10] for a review and comparison of methods). The conventional approach, however, is to make the guess qualitatively by describing the study's limitations. An assessment of the strength of systematic errors, compared with the strength of exposure effects, therefore becomes an exercise in reasoning under uncertainty. Human ability to reason under uncertainty has been well studied and shown to be susceptible to systematic bias resulting in predictable mistakes. A brief review of this literature, focused on situations analogous to epidemiologic inference, suggests that the qualitative approach will frequently fail to safeguard against tendencies to favor exposure effects over systematic errors as an explanation for observed associations. The aforementioned quantitative methods have the potential to safeguard against these failures.

Heuristics and Biases

The Dual Process Model of Cognition

A substantial literature from the field of cognitive science has demonstrated that humans are frequently biased in their judgments about probabilities and at choosing between alternative explanations for observed events[11–14] such as an epidemiologic association. Some cognitive scientists postulate that the mind uses dual processes to solve problems that require such evaluations or choices.[15,16] The first system, labeled the "Associative System," uses patterns to draw inferences. We can think of this system as intuition, although any pejorative connotation of that label should not be applied to the associative system. The second

Submitted 3 May 2006; accepted 19 July 2006.
From Boston University School of Public Health, Boston, MA.
Correspondence: Timothy L. Lash, Associate Professor of Epidemiology, Boston University School of Public Health, Associate Professor of Medicine, Boston University School of Medicine, 715 Albany St., TE3, Boston, MA 02118. E-mail: tlash@bu.edu

ISSN: 1044-3983/07/1801-0067
DOI: 10.1097/01.ede.0000249522.75868.16

system, labeled the "Rule-based System," applies a logical structure to a set of variables to draw inferences. We can think of this system as reason, although the label alone should not connote that this system is superior. The Associative System is not necessarily less capable than the Rule-based System. In fact, skills can migrate from the Rule-based System to the Associative System with experience and the Associative System influences the Rule-based System by affecting the choice of assumptions and axioms used for deduction. The Associative System is in constant action, whereas the Rule-based System is constantly monitoring the Associative System to intervene when necessary. This paradigm should be familiar; we have all said "Wait a minute—let me think," by which we do not mean that we have not yet thought, but that we are not satisfied with the solution our Associative System's thought has delivered. After the chance to implement the Rule-based System, we might say, "On second thought, I have changed my mind," by which we mean that the Rule-based System has overwritten the solution initially delivered by the Associative System.

The process used by the Associative System to reach a solution relies on heuristics. A heuristic reduces the complex problem of assessing probabilities or predicting uncertain values to simpler judgmental operations.[17] An example of a heuristic often encountered in epidemiologic research is the notion that nondifferential misclassification biases an association toward the null. Heuristics often serve us well because their solutions correlate with the truth, but they can sometimes lead to systematic and severe errors, which are called biases.[17] Nondifferential and independent misclassification of a dichotomous exposure leads to the expectation that the measured association will lie between the true association and the null, but many exceptions to the general heuristic exist. Any particular association influenced by nondifferential misclassification may not be biased toward the null.[18] Dependent errors in classification can substantially bias an association away from the null—even if classification errors are nondifferential.[19] Nondifferential misclassification of disease may not lead to any bias in some circumstances.[20] Finally, a true association may not provide stronger evidence against the null hypothesis than the observed association based on the misclassified data—even if the observed association is biased toward the null.[21] Application of the misclassification heuristic without deliberation can lead to errors in an estimate of the strength and direction of the bias,[22] which is also true of general cognitive heuristics.[17]

Cognitive scientists have identified several classes of general heuristics and resulting biases, 3 of which I describe because they may be most relevant to causal inference based on observational epidemiologic results. These heuristics and resulting biases have the following characteristics in common.[23] First, the biases in judgments attributable to the heuristic are systematic and directional; that is, they always act in the same way and in the same direction. Second, they are general and nontransferable; that is, all humans are susceptible to the biases and knowledge of how they act does not immunize us against them. Third, they are independent of intelligence and education; that is, experts make the same mistakes as novices, particularly with a problem that is a little more difficult or a small distance outside of their expertise. Although most studies of these heuristics and biases have been conducted in settings that are not very analogous to causal inference using epidemiologic data, one such study has been conducted and its results corresponded to results elicited in the cognitive science setting.[24] In addition, these heuristics and biases affect evidence-based forecasts of medical doctors, meteorologists, attorneys, financiers, and sports prognosticators.[25] It seems unlikely that epidemiologists would be immune.

Anchoring and Adjustment

The first heuristic relevant to causal inference based on observational epidemiologic results is called "anchoring and adjustment."[17] When asked to estimate an unknown but familiar quantity, respondents use a heuristic strategy to select (or receive) an anchor and then adjust outward from that anchor in the direction of the expected true value. Adjustments are typically insufficient. For example, one might ask the year in which George Washington was elected as the first president of the United States.[26] Most respondents choose the anchor to be 1776, the year that the United States declared independence. Respondents then adjust upward to later years, because they know the U.S. Constitution was not ratified in the same year. The average response equals 1779 and the correct value equals 1788. The predictably insufficient adjustment arises because respondents adjust outward from the anchor until their adjusted estimate enters a range they deem plausible. The true value, more often, lies toward the center of the plausible range. When the anchor is below the true value, like in the year of Washington's first election, the estimate is predictably lower than the true value. Conversely, when the anchor is above the true value, the estimate is predictably higher than the true value. For example, one might ask the temperature at which vodka freezes.[26] Most respondents choose the anchor to be 32°F, the temperature at which water freezes. Respondents adjust downward to lower temperatures, because they know alcohol freezes at a lower temperature than water. The average response equals 1.75°F, and the correct value equals −20°F. Importantly, the anchoring and adjustment heuristic operates in the same manner regardless of whether the anchor is self-generated or externally provided so long as the respondent is aware of the anchor and it is on the same scale as the target.[27]

The anchoring and adjustment heuristic may affect inference from observational epidemiologic results. Consider the point estimate associating an exposure with a disease, derived from a study's results, to be an anchor. Further consider that stakeholders (the investigator, collaborators, readers, and policymakers) may be aware of the direction of an expected systematic error (eg, toward the null). The anchoring and adjustment heuristic suggests that an adjustment to the point estimate to account for the error will be

predictably insufficient. Stakeholders will often adjust the association to account for the error only so far as is plausible, which adjustment will, on average, be insufficient.

Overconfidence

The second bias relevant to causal inference based on observational epidemiologic results is called "overconfidence," which is a systematic error potentially generated by several heuristics. When asked to estimate an unknown but familiar quantity, respondents can be taught to provide a median estimate (the estimate about which they feel it is as likely that the true value is higher as it is that the true value is lower) as well as an interquartile range. The interquartile range is defined by the respondent's estimate of the 25th percentile (the estimate about which they feel it is 75% likely that the true value is higher and 25% likely that the true value is lower) and the respondent's estimate of the 75th percentile. For a well-calibrated respondent, it should be 50% likely that the true value would fall into the interquartile range. For example, one might ask the average annual temperature in Boston, Massachusetts. A respondent might provide a median estimate of 50°F, a 25th percentile estimate of 40°F, and a 75th percentile estimate of 60°F. The true average annual temperature in Boston equals 51.3°F.[28] Were one scoring this respondent's answers, she would receive one point because her interquartile range contains the true value. A second respondent might provide a median estimate of 45°F, a 25th percentile estimate of 40°F, and a 75th percentile estimate of 50°F. Were one scoring this respondent's answers, he or she would receive no point because his or her interquartile range does not contain the true value. Note that the difference in respondents' scores derives more from the narrow width of the second respondent's interquartile range than from the distance of the median estimate from the truth. Were the second respondent's interquartile range as wide as the first respondent's (and centered on the same median estimate), then the second respondent would also have received a positive score. Setting the uncertainty range too narrowly is the hallmark of the overconfidence bias.

In one experiment, a cognitive scientist asked 100 students to answer 10 questions such as the previous question about the average temperature in Boston.[29] For a well-calibrated student, one would expect the true value to lie in the interquartile range for 5 of the 10 questions. Using the binomial distribution to set expectations, one would expect 5 or 6 of the 100 students to give answers such that 8, 9, or 10 of the true values fell into their interquartile ranges. None of the students had scores of 8, 9, or 10. One would also expect 5 or 6 of the 100 students to give answers such that 2, 1, or 0 of the true values fell into their interquartile ranges. Thirty-five of the students had scores of 2, 1, or 0. The skew toward low scores arises because respondents provide too narrow a range of uncertainty, so the true value lies outside the interquartile range much more often than it lies inside it. The overconfidence bias acts in the same way when respondents are asked to give extreme percentiles such as the 1st and 99th percentiles[29] is most pronounced when tasks are most difficult[30] has been observed to act in many different populations and cultures[31] and does not depend strongly on how well respondents estimate the median.[29] In fact, the discrepancy between correctness of response and overconfidence increases with the knowledge of the respondent. That is, when a response requires considerable reasoning or specialized knowledge, the answers of experts are more accurate (ie, more often correct or nearer to correct) than the answers of novices. However, the experts' overconfidence—compared with novices—increases faster than their accuracy.[32]

Overconfidence may affect inference from observational epidemiologic results. Consider the conventional frequentist confidence interval about a point estimate associating an exposure with a disease, derived from a study's results, to be an uncertainty range analogous to the interquartile range described here. Further consider that stakeholders may be aware that the interval fails to account for uncertainty beyond random error and so should be considered a minimum description of the true uncertainty. The overconfidence bias suggests that an intuitive inflation of the confidence interval to account for sources of uncertainty aside from random error will be predictably insufficient.

Failure to Account for the Base Rate

The final bias relevant to causal inference based on observational epidemiologic results is called "failure to account for the base rate," a result of the representativeness heuristic.[33] When asked to estimate the probability of an event on the basis of the base rate frequency of the event in a relevant reference population and specific evidence about the case at hand, respondents systematically focus on the specific evidence and largely ignore the base rate information.[34] For example, 60 medical students were asked[35]: If a test to detect a disease whose prevalence is one in 1000 has a false- positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease assuming you know nothing about the person's symptoms or signs?

Almost half of the respondents answered 95%, which takes account of only the specific evidence (the patient's positive test) and completely ignores the base rate information (the prevalence of the disease in the population). Eleven students gave the correct response (2%). Failure to account for the base rate does not derive solely from innumeracy.[36] Rather, the specific evidence is concrete and emotionally interesting, thereby more readily inspiring a mental script to explain its relevance. Base rate information is abstract and emotionally uninteresting, so less likely to inspire an explanatory script that contradicts the specific evidence. In some experiments, framing the problem in terms of frequencies rather than probabilities reduces susceptibility to failure to account for the base rate.[37]

Failure to account for the base rate may affect inference from observational epidemiologic results. Consider a conventional epidemiologic result, comprised of a point estimate associating an exposure with a disease and its frequentist confidence interval, to be specific evidence about a hypothesis that the exposure causes the disease. Further consider that stakeholders have devoted considerable effort to gener-

ating and understanding the research results. The "failure to account for the base rate" bias suggests that stakeholders are not likely to adequately account for the base rate of "true" hypotheses studied by epidemiologists despite exhortations to use base rate information in epidemiologic inference.[38–40]

DISCUSSION

Epidemiologists are not alone among scientists in their susceptibility to the systematic errors in inference engendered by the heuristics and biases described here. For example, a review of measurements of physical constants reported consistent underestimation of uncertainty.[41] Measurements of the speed of light overestimated the currently accepted value from 1876 to 1902 and then underestimated it from 1905 to 1950. This pattern prompted one investigator to hypothesize a linear trend in the speed of light as a function of time and a second investigator to hypothesize a sinusoidal relation. In reaction, Birge[42] adjusted a set of measurements for systematic errors, produced corrected values and intervals that overstated—rather than understated—the uncertainty, and concluded that the speed of light was constant.[42] Henrion and Fischoff[41] attribute the consistent underassessment of uncertainty in measurements of physical constants to investigators using the standard error as the full expression of the uncertainty regarding their measurements to the impact on their inferences of heuristics and biases such as those described here and to "real-world" pressures that discourage a candid expression of total uncertainty. These same 3 forces likely affect inference from observational epidemiology studies as well.

Henrion and Fischoff[41] recommend 3 solutions. First, those who measure physical constants should strive to account for systematic errors in their quantitative assessments of uncertainty. Second, with an awareness of the cognitive literature, those who measure physical constants should temper their inference by subjecting it to tests that counter the tendencies imposed by the heuristics and biases. For example, overconfidence arises in part from a natural tendency to overweigh confirming evidence and to underweigh disconfirming evidence. Forcing oneself to write down hypotheses and evidence that counter the preferred (ie, causal) hypothesis can reduce overconfidence in that hypothesis. Last, students should be taught how to obtain better measurements, including how to better account for all sources of uncertainty and how to counter the role of heuristics and biases in reaching an inference. These same recommendations would well serve those who measure epidemiologic associations.

Reducing our enthusiasm about the results we generate may seem like a suitable alternative. In cognitive sciences literature, this approach is called debiasing and sorts into 3 categories.[43] These categories are: resistance—"a mental operation that attempts to prevent a stimulus from having an adverse effect," remediation—"a mental operation that attempts to undo the damage done by the stimulus," and behavior control—"an attempt to prevent the stimulus from influencing behavior." These strategies are ineffective solutions to tempering the impact of the heuristics and biases described here.[43]

Nonetheless, epidemiologists are taught to rely on debiasing when making inference. We are told to interpret our results carefully and to claim causation only with trepidation. Consider, for example, the disparity between randomized[44] and nonrandomized[45] studies of the association between hormone replacement therapy and cardiovascular disease, one of the most recent and high-profile examples of hypotheses supposedly established by observational epidemiologic research and subsequently reversed or discounted. Three commentators offered advice tantamount to a warning to "be careful out there." One wrote that we should be "reasonably cautious in the interpretation of our observations,"[46] the second wrote "we must remain vigilant and recognize the limitations of research designs that do not control unobserved effects,"[47] and the third wrote "future challenges include continued rigorous attention to the pitfalls of confounding in observational studies."[48] Similar warnings are easy to find in classroom lecture notes and textbooks.

The reality is that such trepidation, even if implemented, is ineffective. Just as humans overstate their certainty about uncertain events in the future, we also overstate the certainty with which we believe that uncertain events could have been predicted with the data that were available in advance had they been more carefully examined. The tendency to overstate retrospectively our own predictive ability is colloquially known as "20–20 hindsight." Cognitive scientists, however, label the tendency "creeping determinism."[11]

Creeping determinism can impair one's ability to judge the past or to learn from it. It seems that when a result such as the trial of hormone therapy becomes available, we immediately seek to make sense of the result by integrating it into what we already know about the subject. In this example, the trial result made sense only with the conclusion that the nonrandomized studies must have been affected by unmeasured confounders, selection forces, and measurement errors, and that the previous consensus must have been held only because of poor vigilance against systematic errors that act on nonrandomized studies. With this reinterpretation, the trial results seem an inevitable outcome of the reinterpreted situation. Making sense of the past consensus is so natural that we are unaware of the impact that the outcome knowledge (the trial result) has had on the reinterpretation.[49] Therefore, merely warning people about the dangers apparent in hindsight such as the recommendations for heightened vigilance quoted previously has little effect on future problems of the same sort.[11] A more effective strategy is to appreciate the uncertainty surrounding the reinterpreted situation in its original form. For example, several epidemiologists had questioned the preventive relation between hormone replacement therapy and cardiovascular disease before the trials,[50–53] so the uncertainty engendered by their original criticisms should now receive due attention. Better still is to appreciate the uncertainty in current problems, applying lessons from the past such as the concurrent questions raised by these epidemiologists, to avoid similar problems in the future.

In fact, cataloging the uncertainty surrounding events is the one method of removing bias that reliably reduces over-

confidence. Simply weighing whether a hypothesis is true—equivalent to the "vigilance" recommended by editorialists on the putative cardioprotective effect of hormone replacement therapy—actually increases belief of the validity of the hypothesis because a person focuses more on explanations as to why it could be true than why it could be false.[43] To remove bias, a person must consider the opposite. That is, one must imagine alternative hypotheses, which should illuminate the causal hypothesis as only one in a set of competing explanations for the observed association. Viewing infrequent events in their set structure improves probability judgments.[54] These general recommendations have been suggested to epidemiologists as constructive methods to reveal uncertainties.[55]

Conventional observational research treats nonrandomized data as if it were randomized in the analysis phase. In the inference phase, investigators usually offer a qualitative discussion of limitations, which are too often discounted without quantification as important threats to validity. Alternatives that quantify threats to validity do exist but have been infrequently adopted. Instead, epidemiologists are warned to make inference with trepidation because of the potential for systematic error. This paradigm corresponds quite well with circumstances known to be ripe for the impact of the heuristics and biases described here. In the presence of sparse evidence and a low base rate of true hypotheses, those who assess the probability of an event—such as the truth of the hypothesis—are overconfident, on average.[25]

Epidemiologists could explicitly specify alternatives to the causal hypothesis and quantify the uncertainty about the causal association induced by each alternative hypothesis. This alternative paradigm removes the guesswork about sizes of systematic errors, compared with the size of exposure effects, from the Associative System and places it under the purview of the Rule-based System. The second paradigm prompts investigators to list explanations for results that counter their preferred hypothesis and requires that they incorporate these hypotheses into their assessments of uncertainty. Were such a list and quantification to become the norm, then confidence in the causal explanation would be appropriately reduced compared with the status quo, even if the list was not exhaustive or the quantification was not accurate. The result would be a more complete description of total uncertainty and an effective counter to the impact of the heuristics and biases described here.

ACKNOWLEDGMENT

The author thanks Steven A. Sloman of Brown University for his helpful comments on a draft of the manuscript.

REFERENCES

1. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1:421–429.
2. Espeland MA, Hui SL. A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*. 1987;43:1001–1012.
3. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology.* Boca Raton: Chapman and Hall/CRC; 2003.
4. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error, in main study/validation study designs. *J Am Stat Assoc*. 2000;95:51–61.
5. Little RJA, Rubin DB. *Statistical Analysis With Missing Data,* 2nd ed. New York: Wiley; 2002.
6. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.
7. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003;14:451–458.
8. Phillips CV. Quantifying and reporting uncertainty from systematic errors. *Epidemiology*. 2003;14:459–466.
9. Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol*. 2004;33:1389–1397.
10. Greenland S. Multiple-bias modeling for analysis of observational data (with discussion). *Royal Stat Soc*. 2005;168:267–306.
11. Piattelli-Palmarini M. *Inevitable Illusions.* New York: Wiley; 1994.
12. Gilovich T. *How We Know What Isn't So.* New York: The Free Press; 1991.
13. Kahneman D, Slovic P, Tversky A. *Judgment Under Uncertainty: Heuristics and Biases.* New York: Cambridge University Press; 1982.
14. Gilovich T, Griffin D, Kahneman D. *Heuristics and Biases: The Psychology of Intuitive Judgment.* New York: Cambridge University Press; 2002.
15. Kahneman D, Frederick S. Representativeness revisited: Attribute Substitution in intuitive judgment. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:49–81.
16. Sloman SA. Two systems of reasoning. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:379–396.
17. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982:3–22.
18. Jurek A, Greenland S, Maldonado G, et al. Proper interpretation of nondifferential misclassification effects: expectations versus observations. *Int J Epidemiol*. 2005;34:680–687.
19. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*. 1992;3:210–215.
20. Brenner H, Savitz DA. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case–control studies. *Am J Epidemiol*. 1990;132:181–192.
21. Greenland S, Gustafson P. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *Am J Epidemiol*. 2006;164:63–68.
22. Lash TL, Fink AK. Neighborhood environment and loss of physical function in older adults: evidence from the Alameda County study [Letter]. *Am J Epidemiol*. 2003;157:472–473.
23. Piattelli-Palmarina M. How to emerge from the tunnel of pessimism. In: Piattelli-Palmarini M, ed. *Inevitable Illusions*. New York: Wiley; 1994: 139–145.
24. Holman CDJ, Arnold-Reed DE, de Klerk N, et al. A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. *Epidemiology*. 2001;12:246–255.
25. Koehler DJ, Brenner L, Griffin D. The calibration of expert judgment: Heuristics and biases beyond the laboratory. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:686–715.
26. Epley N, Gilovich T. Putting adjustment back in the anchoring and adjustment heuristic. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:139–149.
27. Chapman GB, Johnson EJ. Incorporating the irrelevant: Anchors in judgments of belief and value. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:120–138.
28. www.climate-zone.com/climate/united-states/massachusetts/boston. Accessed 24 January 2006.
29. Alpert M, Raiffa H. A progress report on the training of probability assessors. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982:294–305.

30. Lichtenstein S, Fischoff B, Phillips LD. Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982:306–334.
31. Yates JF, Lee JW, Sieck WR, et al. Probability judgment across cultures. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:271–291.
32. Piattelli-Palmarini M. *Inevitable Illusions.* New York: Wiley; 1994; 119.
33. Lagnado DA, Sloman SA. Inside and outside probability judgment. In: Koehler DJ, Harvey N, eds. *Blackwell Handbook of Judgment and Decision Making*. Oxford, UK: Blackwell Publishing; 2004:157–176.
34. Tversky A, Kahneman D. Evidential impact of base-rates. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982:153–162.
35. Casscells W, Schoenberger A, Grayboys T. Interpretation by physicians of clinical laboratory results. *N Engl J Med*. 1978;299:999–1000.
36. Nisbett RE, Borgida E, Crandall R, et al. Popular induction: Information is not necessarily informative. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982:101–116.
37. Gilovich T, Griffin D. Introduction—heuristics and biases: then and now. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:1–18.
38. Poole C. Low P-values or narrow confidence intervals. Which are more durable? *Epidemiology*. 2001;12:291–294.
39. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004;96:434–442.
40. Greenland S, Robins J. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*. 1991;2:244–251.
41. Henrion M, Fischoff B. Assessing uncertainty in physical constants. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:666–677.
42. Birge RT. The general physical constants: as of August 1941 with details on the velocity of light only. *Reports on Progress in Physics*. 1941;8:90–134.
43. Wilson TD, Centerbar DB, Brekke N. Mental contamination and the debiasing problem. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press; 2002:185–200.
44. Writing group for the woman'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;288:321–333.
45. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med*. 1991;20:47–63.
46. Michels KB. Hormone replacement therapy in epidemiologic studies and randomized clinical trials—are we checkmate? *Epidemiology*. 2003;14: 3–5.
47. Piantadosi S. Larger lessons from the Women's Health Initiative. *Epidemiology*. 2003;14:6–7.
48. Whittemore AS, McGuire V. Observational studies and randomized trials of hormone replacement therapy: what can we learn from them? *Epidemiology*. 2003;14:8–10.
49. Prentice RL, Langer R, Stefanick ML, et al, for the Women's Health Initiative Investigators. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative Clinical Trial. *Am J Epidemiol.* 2005;162:404–414.
50. Petitti DB, Perlman JA, Sidney S. Postmenopausal estrogen use and heart disease. *N Engl J Med*. 1986;315:131–132.
51. Ward FM, Posthuma M, Westendorp RGJ, et al. Cardioprotective effect of hormone replacement therapy in postmenopausal women: is the evidence biased? *BMJ*. 1994;308:1268–1269.
52. Vandenbroucke JP. How much of the cardioprotective effect of postmenopausal estrogens is real? *Epidemiology*. 1995;6:207–208.
53. Sturgeon SR, Schairer C, Brinton LA, et al. Evidence of a healthy estrogen survivor effect. *Epidemiology*. 1995;6:227–231.
54. Sloman SA, Over D, Slovak L, et al. Frequency illusions and other fallacies. *Organ Behav Hum Decis Process*. 2003;91:296–309.
55. Savitz DA. *Interpreting Epidemiologic Evidence.* New York: Oxford University Press;2003;44: