Living with Statistics in Observational Research

Sander Greenland[a,b] *and Charles Poole*[c]

We thank Andrew Gelman for his comments[1] on our article.[2] We hope our response will clarify areas of agreement and disagreement. We also take the opportunity to address some larger issues about apparent disconnections: the disconnection between conventional statistical models and the realities of health and social-science research, and the disconnection between what methodologists seem to say one should do and what everyone, including methodologists, do (or, rather, do not do) in reality.

BAYESIANS AND FREQUENTISTS VERSUS SCIENTISTS

It seems to us that Bayesian viewpoints and methods have a greater ability to reflect observational research realities than do conventional frequentist methods,[3] although relatively unconventional frequentist methods have been developed to address these concerns.[4] That is largely because conventional methods start from models of ideal experiments and elaborate them, always ending with a model that assumes "no confounding" or "ignorability" or the like, which is operationally equivalent to claiming that our data were produced by some kind of intricately designed randomized experiment. These approaches never confront the fact that (by definition) observational research is about situations in which no experiment has been conducted.

In recommending Bayesian perspectives, we emphasize that we are neither Bayesians nor frequentists. Nor do we recommend that anyone become either of them, or anything other than a good rational scientist open to using whatever tools work well for the immediate task.[5,6] Pure Bayesian methodology is clearly insufficient if frequency performance is of concern (as it often is) and has other limits as well.[7,8] Many statisticians who include Bayesian methods prominently in their toolkit emphasize the need for model checks, including frequentist devices such as *P* values.[9–13] Simply put, a statistician claiming that one statistical philosophy (whether Bayesian, frequentist, or another) is preferred for all analyses would be like a carpenter claiming that screws are always better than nails for joining wood.

More generally and relevantly for epidemiology, in the preface to his remarkable 1978 book on statistical modeling, *Specification Searches*[14] (out of print but available free online at the Anderson Project website), the econometrician Edward Leamer wrote:

> I am confident that the Bayesian approach helps us understand nonexperimental inference. It helps also to avoid certain errors. But I do not think it can truly solve all the problems. Nor do I foresee developments on the horizon that will make any mathematical theory of inference fully applicable. For better or for worse, real inference will remain a highly complicated, poorly understood phenomenon.[15]

A generation earlier, Jerome Cornfield, perhaps the first Bayesian epidemiologic statistician, voiced similar reservations about statistical developments:

From the [a]Department of Epidemiology and the [b]Department of Statistics, University of California, Los Angeles, CA; and the [c]Department of Epidemiology, University of North Carolina, Chapel Hill, NC.

Editors' note: *Related articles appear on pages 62 and 69.*

Correspondence: Sander Greenland, Department of Epidemiology and Department of Statistics, University of North California at Los Angeles, Topanga, CA 90290. E-mail: lesdomes@ucla.edu.

ISSN: 1044-3983/13/2401-0073
DOI: 10.1097/EDE.0b013e3182785a49

> Enthusiasm for the newer statistical tools, while deserved, should be tempered with recognition of the importance of insight, imagination, and intimate knowledge of one's field. The statistician functions as a devil's advocate against the admission of new evidence, and in this capacity, has an important influence on the quality and cogency of the evidence submitted. The investigator must pay close attention to this advocate but it is his and not the advocate's responsibility to decide when he must stop listening.[16]

It is sobering to realize that Mantel-Haenszel methods were the newest statistical tools in epidemiology when Cornfield wrote this, in the era of computers fed with punch cards. Modeling and computing capabilities have grown by many orders of magnitude since then, yet the inferential capabilities of statistics (whether Bayesian, frequentist, or other) are still profoundly limited. We thus think the above cautions apply unchanged today.

In its purest form, frequentism is the use of methods based solely on "long-run" frequency performance, which consigns to informal (and hence easily neglected) status other operational criteria for adequacy, such as coherence between contextual (background or "prior") information and the data model. Like any other pure approach, pure frequentism is a doctrinaire philosophy that fails in rather glaring ways. The ever-shifting environments studied by health and social sciences ensure that long-run performance under the current analysis model is almost never a directly relevant validity criterion; as Keynes[17] said, "In the long run we are all dead." This problem is all the more acute in fields like epidemiology in which changes over a subject's life may have effects that can drown out the effects we are trying to study.

Another reason is that we never know what the long run holds in store (apart from death and taxes), yet conventional statistics use models about the long run that in essence claim we do know what will happen, at least in broad structural outline. A benefit of the subjective-Bayesian perspective is that it treats these models as nothing more than conjectures about the long run rather than objective reality. Viewing frequentist analyses of observational data in this harsh light reveals that they depend on implausible assumptions that too often turn out to be so wrong as to deliver unreliable inferences. Objective Bayesian methods use the same data models as frequentists, and so are subject to parallel criticism. Subjective Bayesians usually use the same models, but without elevating the models or their outputs to the status of objective facts.

The heroic subjective-Bayesian advocate Dennis Lindley (1982) put the problem in a way any competent epidemiologist should appreciate:

> Because scientific measurements typically contain unknown and undetected biases, precision can increase without limit but accuracy cannot. Statisticians, with their emphasis on standard errors that ignore the bias, have confused the issue in some scientific experimentation because the error they quote is substantially less than the true error.[18]

His complaint applies with even greater force to observational research.[14,19] Gelman's discussion of Bem[20] on extrasensory perception (ESP) echoes this complaint, as the problems Gelman raises (selection bias, misclassification) make every conventional statistic misleading—including confidence intervals and *P* values.

Despite early recognition of its profound shortcomings, frequentism achieved something of a dictatorial monopoly in the sciences in the mid-20th century (Cornfield being a notable exception in epidemiology). Claims that its enduring popularity must have been because the methods work well in practic e[21p.319,22p.2] strike us as much like claiming that trepanation and bloodletting must have worked well because they were taught and practiced for centuries by the most respected experts. For example, bleeding as a treatment for yellow fever was defended (in part on proto-epidemiologic grounds) by Benjamin Rush in the late 18th century,[23] yet few today would doubt that bloodletting only worsened morbidity and mortality. We see many parallels in modern statistical practice. Among them, the frequentist hegemony has produced an epidemic of statistical testing that in turn has led to gross distortions of data reporting.[24–26]

SIGNIFICANCE TESTS: TOO EASY TO COMPUTE AND TOO HARD TO INTERPRET CORRECTLY

Going beyond historical divides, we have argued that the usual criteria (be they Bayesian, frequentist, likelihoodist, or other) for "working well" are at best insufficient and at worst misleading for observational research. There is one criterion, however, that frequentist statistics fail miserably: ease of correct interpretation.

Significance tests and *P* values for null hypotheses are exceptionally easy to compute, which is a major reason they were adopted so widely and so early in the era before digital computers. But significance tests and null *P* values are exceptionally difficult to interpret correctly, even in the hands of statistics professors discussing randomized trials.[27,28] *P* values and confidence intervals become even harder to interpret correctly in observational studies, especially in light of the reality that a single study will seldom form the entire basis for any decision about an effect, and no study will be free of biases (which render ludicrous any claims that we can precisely distinguish "significant" from "nonsignificant"). We have thus attempted to promote more defensible non-Bayesian interpretations that do not employ or imply any magical cutoff for decisions or inference.[26,27,29] As with confidence intervals, however, these interpretations remain problematic owing to the presence of bias in nearly all studies (including randomized trials) and by the hazards of inference from single studies.

CATEGORIZATION OF *P* VALUES GUARANTEES MISINTERPRETATION

We laud Gelman for calling attention to the instability of *P* values and hence of "significance" and "nonsignificance"

declarations,[30] although, in at least some work,[31] Gelman seems to use the conventional 0.05 "significance" cutoff. Here he states that a trichotomy not involving that value is typical:

> In theory the *P* value is a continuous measure of evidence, but in practice it is typically trichotomized approximately into *strong evidence*, *weak evidence*, and *no evidence* (these can also be labeled highly significant, marginally significant, and not statistically significant at conventional levels), with cutoffs roughly at $P = 0.01$ and 0.10.[1]

Although others also mention this trichotomy,[32] we have not seen it used in health-science studies. What we see most often instead is the well-documented "cliff effect" at 0.05.[33,34] Using the 0.05 dichotomy, an effect is considered "demonstrated" if $P = 0.04$,[35] but a study has "failed" if $P = 0.052$.[36] Researchers believe they can "establish" effects with $P = 0.03$,[37] but interpret results as "showing" that risk "does not increase" when $P = 0.09$.[38] To the extent that we see a third category used, it is in the region from exactly 0.05 to approximately 0.10. *P* values in this range are occasionally called "borderline"[39] or, in a regrettable choice of words,[40] a "trend."[41]

No trichotomy for *P* values would resolve the issues Gelman raises. Consider his example of two independent experiments with estimates (standard errors) of 25 (10) and 10 (10). We lament, as he does, the tragedy that many would call these two experiments contradictory, or say that the second did not "replicate" the first, because the null hypothesis was rejected in the first test (Wald $P < 0.05$) but not in the second ($P > 0.05$). These are wrong conclusions, even within the null hypothesis testing framework, because the null hypothesis of no difference between the two experiments is not rejected ($P > 0.05$). Using a significance testing approach and an evidentiary trichotomy leaves an apparent inconsistency among these results: The first experiment provides "strong evidence" against the null ($P \approx 0.01$) and the second provides "no evidence" against null ($P \approx 0.3$); yet the comparison of the two provides "no evidence" that they differ ($P \approx 0.3$)!

Still, we think a trichotomy is preferable to a dichotomy because a third, indeterminate option between rejection (significance) and nonrejection (nonsignificance) might reduce some of the dichotomy's damage, such as publication bias. Suspending firm decisions (ie, interpreting results with extra caution) pending examination of other evidence is usually a good option to leave on the table. If one were to adopt a trichotomous approach, the boundaries of the middle range of no decision or inference should depend on the context, including the alternatives in play and the costs of errors—just as Neyman[42] advised for dichotomous decision cutoffs (alpha levels).

An even healthier way to interpret *P* values, however, would be as a continuous measure, in the style advocated (if not consistently practiced) by Fisher.[32,40,43–46] In doing so, we would see that the interpretation of $P > 0.10$ as "no association," "no evidence of an association," or "no evidence" is simply wrong. In Fisher's interpretation, every *P* value corresponds to a degree of evidence against the entire statistical model (the set of assumptions) in which the tested hypothesis is embedded.[47,48] For example, a null *P* value for the coefficient in a causal (structural) model might be very small because the effect is very large or because some other aspect of the model is very wrong (eg, owing to validity problems). Only a *P* value of exactly 1 represents no evidence against the model. All *P* values below 1 represent evidence against the model, and smaller *P* values represent more evidence, with a *P* value of exactly 0 representing logical contradiction of the model.

Most usefully, for epidemiology and other fields in which the entire model surrounding the test is questionable, this shift from interpreting the *P* value as a test of a hypothesis to a test of its embedding model applies whether or not biases are present.[47,48] An essential caution in this use is that a large *P* value does not mean the model should be taken as correct; rather, it means that the data alone do not provide enough information for the test to detect model defects, which still may be large in practical terms. Another caution is that there are arguments for not using *P* values as measures of evidence[10,49,50] as opposed to tests of fit, for example, that *P* values are poorly scaled for measuring evidence when compared to likelihood ratios and related information measures.

THE IMPORTANCE OF ALTERNATIVES AND THE DECEPTIVENESS OF POWER

Alternative hypotheses and a comparative concept of statistical evidence are central to both Neyman-Pearson testing theory and likelihood theory, and they are easily accommodated within Fisher's continuous view of significance tests. For example, a *P* value of 0.20 for relative risk (RR) = 1 may be considered weak evidence against the null, but if the *P* value for a proposed alternative of RR = 1.5 is 0.90, the evidence against RR = 1.5 is even weaker and thus the data alone provide no basis for preferring or inferring RR = 1 over RR = 1.5. The same point can be seen more directly by computing the likelihood ratio comparing RR = 1 to RR = 1.5.[27,49]

One misguided and complicated attempt to consider alternatives examines post hoc power (power calculations on observed data), which Gelman and Weakliem[31] use in their sex-ratio analysis. Post hoc power is completely dependent on the chosen significance cutpoint, and exhibits highly counterintuitive behavior.[51] For example, a test of the null can be "nonsignificant" ($P > 0.05$) and have "high power" (>80%) against a specified alternative, yet at the same time that alternative can have a higher *P* value and likelihood than the null.[52] It is far easier to understand comparisons of *P* values, likelihood ratios, or posterior probabilities, and for these comparisons no cutpoint is needed. Indeed, Gelman[1] and Gelman and Stern[30] point out the statistical unreliability of cutpoint-driven interpretation of *P* values; thus, Gelman and Weakliem's[30] usage of power[31] seems to illustrate only how hard it is to break the "significance" habit.

WHY WE TURNED TO BAYESIAN INTERPRETATIONS OF *P* VALUES

Although presenting and thinking of *P* values on a continuum may avert some of their worst misinterpretations, the correct frequentist interpretation (which Gelman[1] reviews) is obtuse, even from a continuous viewpoint. It is so obtuse that *P* is usually misinterpreted as the probability that chance alone produced the observed association. For example, Oleckno[53 p.182] states that the *P* value "measures the probability that the difference is due to sampling error"; Marciante et al[54] state that they "conducted a permutation test to estimate the probability of a chance finding"; and Harris and Taylor[55] state that "the *P* value gives the probability of any observed differences having happened by chance." Unfortunately for these misinterpretations, the probability that sampling error or random error or chance alone produced a difference or an association or a "finding" is logically identical to the probability that the null hypothesis is true and there is no bias, which is a Bayesian posterior probability.[27] In practice, the null *P* value is rarely even close to this posterior probability.[28] Thus, it seems valuable to describe the posterior probabilities that the null *P* value actually does approximate or bound, our paper's main purpose.

Consider again Gelman's example of two independent experiments with estimates (standard error) of 25 (10), with null *P* value $P_0 = 0.012$, and 10 (10), $P_0 = 0.32$. We may interpret these as saying that, if perfect, the first trial by itself should leave us with a posterior probability of at least $P_0/2 = 0.6\%$ that the effect underlying the first trial is negative; whereas, if perfect, the second trial by itself should leave us with at least $P_0/2 = 16\%$ posterior probability that the effect underlying the second trial is negative. The one-sided *P* value for their difference, $0.289/2 \approx 0.14$, says that, if perfect, these trials by themselves should leave us with at least 14% probability that the effect underlying the first trial is actually smaller than the effect underlying the second trial, opposite of what the estimates suggest. Unlike the conventional "significance" statements or evidence categorizations, these correct Bayesian interpretations strike us as intuitively consistent with one another.

Another interesting divergence of the bounding interpretation of *P* values from their usual "significance" interpretation is that the smaller a *P* value, the less informative it is as a lower bound. For example, being left with at least $P_0/2 = 0.6\%$ probability our estimate is on the wrong side of the null says much less than being left with at least $P_0/2 = 16\%$ probability our estimate is on the wrong side of the null. Thus, it is the larger value of $P_0/2$ that is the more informative of the two numbers. We think this property neatly answers Gelman's concern that the one-sided null *P* value $P_0/2$ seems too extreme numerically as a posterior probability: $P_0/2$ seems extreme only if one fails to recognize that it is a lower bound, which means the smaller it is the less information it conveys—just the opposite of the way most everyone perceives *P* values.

For related reasons, it is not generally true that "analyses assuming broader priors will systematically overstate the probabilities of very large effects."[1] This statement is correct only if one's prior distribution is concentrated near the null, as is Gelman's prior in all his examples, which suggests to us that Gelman (like many) approaches analyses with such a prior as his implicit default. But interpreting $P_0/2$ as a directional posterior probability is predicated on the assumption that we have little background information, which is the reality for many epidemiologic studies (such as those concerned with side effects of relatively new drugs or rare environmental exposures).

With ESP and with sex ratios, a strong null-centered prior is an opinion derivable from many studies conducted over many generations. Like Gelman, we would not want to mistake $P_0/2$ as our actual posterior probability in these examples. With such strong and well-founded priors, $P_0/2$ would be uselessly low (if not misleading), even as a lower bound. This is another way of saying that strong and well-founded priors create precisely the setting in which a given study should not be used for inference by itself.

Nonetheless, we fear that Gelman overlooked a hazard complementary to ignoring prior information: the use (often implicit) of strong priors that are not well founded. We would point two cautionary examples: the famed Bayesian, Sir Harold Jeffreys (a geophysicist as well as statistician), who insisted with high certainty that continents do not drift, and the even more famous Sir Ronald Fisher (a geneticist as well as statistician), who insisted that the relation of smoking to lung cancer could be easily explained away by confounding. Such cautionary tales of great scientists being misled by their own strong but poorly founded priors has led to a theme in Bayesian methodology: Whatever our prior opinion and its foundation, we still need reference analyses with weakly informative priors to alert us to how much our prior probabilities are driving our posterior probabilities. In this role, $P_0/2$ is far from ideal (because, like all other conventional statistics, it ignores uncontrolled biases), but it is better than no reference at all.

SPIKING SPIKED PRIORS

We agree with Gelman[1] that "a Bayesian interpretation based on a spike-and-slab model makes little sense in applied contexts in epidemiology, political science, and other fields in which true effects are typically nonzero"; see also Gelman et al.[56] As we would put it, in health and social sciences there is rarely any positive scientific evidence that the null is exactly true and only a few specialties (eg, genomics) with remotely credible mechanistic arguments for the exact null. Cox[57] opined similarly that in many studies "there may be no reason for expecting the effect to be null. The issue tends more to be whether the direction of an effect has been reasonably firmly established and whether the magnitude of any effect is such as to make it of public health or clinical importance," which leads directly to using one-sided in place of two-sided

P values. Thus, for most applications, there is no scientific basis for placing a point probability mass at the null or any other point.

Our stand against spikes directly contradicts a good portion of the Bayesian literature, where null spikes are used too freely to represent the belief that a parameter "differs negligibly" from the null. In many settings we see, even a tightly concentrated probability near the null has no basis in genuine evidence. Many scientists and statisticians exhibit quite a bit of irrational prejudice in favor of the null based on faith in oversimplified physical models; Shermer[58] is a vivid example involving cell phones and cancer (see the Greenland[59] chapter for a discussion). This null prejudice also arises more subtly from confusion of decision rules with inference rules, and from adoption of simplicity or parsimony as a metaphysical principle rather than as an effective heuristic (see, for example, the writings of Kelly[60] for a critical analysis of the distinction).

Cultural norms vary among research areas on this question. In psychological research, for instance, many hold that the null hypothesis is almost never true.[61–65] We may be highly certain that any effect present is small enough so that it would make sense to behave as if the null were true until presented with sufficient evidence otherwise (a practice both Fisher and Neyman recommended); this is a heuristic use of parsimony. But a prior that a hypothesis is (for now) a useful approximation to the truth can lead to results quite different from using a spiked prior (which presumes there is evidence that the tested hypothesis is exactly true).[66] When there is no such evidence, a spike represents an unscientific faith in, or commitment to, the null, with no empirical foundation in most health and social-science applications.[67] The only remaining rationale for spikes is then as crude devices for model or variable selection, but in that role they have been surpassed by modern "hard" shrinkage and resampling techniques as found in the statistical-learning literature.[68,69]

CONCLUSIONS

Faced with the enduring ubiquity of *P* values, statistical tests, confidence intervals, and their misinterpretations, we have countered with correct interpretations.[27] The predominant misinterpretations are Bayesian in form, by researchers trying to answer Bayesian questions with frequentist statistics ill-suited to the task. Thus, the correct interpretations of one-sided *P* values we have most recently reviewed are Bayesian. These interpretations reveal a deep flaw in the common perception that small *P* values are more informative than large *P* values, showing that from at least one Bayesian perspective just the opposite is true.

Like the usual frequentist interpretations, these Bayesian interpretations leave the data model unchallenged, and assume that prior information about the model parameters is weak as compared with the information in the likelihood or estimating equations used to compute the *P* values. This assumption is made explicit by us and by Casella and Berger[70,71] but is implicit in any correct Bayesian interpretation of conventional frequentist statistics. Common examples include the interpretation of a 95% confidence interval as "a range of values either side of the estimate between which we can be 95% sure that the true value lies"[72] or within which "we are 95% confident" that it lies.[54,73–75]

Do we need the reinterpretations of *P* values we described? We think yes, as long as some insist on presenting "significance" statements that are nothing more than dichotomized *P* value reports (as Gelman[1] does in his examples), and as long as others insist on misinterpreting a null *P* value as the probability of the null or the probability that chance alone produced the observations. Upon encountering the latter misinterpretations, editors, reviewers, and readers need to contrast them against correct Bayesian interpretations. To reject those correct interpretations will only help perpetuate the misinterpretations that dominate the scientific literature.

When we are the analysts and wish to provide a more realistic analysis, a *P* value is simply a reference point: It gives us a bound on a posterior probability under the conventional assumptions, for contrast to the corresponding posterior probability when we make more realistic assumptions. If all epidemiologic analyses employed realistic, reasonable assumptions and models, there might be no need for this reference point; traditional statistical tests, *P* values, and confidence intervals could then be confined to those randomized experiments in which their assumptions were met (and where they would still need to be interpreted correctly). Because we do not see this happening in our lifetimes, we advise epidemiologists to study correct interpretations of these statistics, if only to immunize themselves and caution their collaborators against some of the more damaging misinterpretations and misuses.

REFERENCES

1. Gelman A. *P* values and statistical practice. *Epidemiology*. 2013;24:69–72.
2. Greenland S, Poole C. Living with *P*-values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*. 2012;24:62–68.
3. Greenland S. Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. *Int J Epidemiol*. 2009;38:1662–1673, corrigendum (2010) *International Journal of Epidemiology* 39:1116.
4. Vansteelandt S, Goetghebeur E, Kenward MG, Molenberghs G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*. 2006;16:953–980.
5. Neyman J. Statistics: servant of all sciences. *Science*. 1955;122:401–406.
6. Savitz DA. The alternative to epidemiologic theory: whatever works. *Epidemiology*. 1997;8:210–212.
7. Pearl J. Bayesianism and causality, or why I am only half Bayesian. In: Corfield D, Williamson J, eds. *Foundations of Bayesianism. Kluwer Applied Logic Series*. Dordrecht,The Netherlands: Kluwer Academic Publishers; 2001;24:19–36.
8. Kelly K, Glymour C. Why probability does not capture the logic of scientific justification. In: Hitchcock C, ed. *Contemporary Debates in the Philosophy of Science*. London, UK: Blackwell; 2004.
9. Box GEP. Sampling and Bayes inference in scientific modeling and robustness. *J R Stat Soc Ser A*. 1980;143:383–430.
10. Good IJ. *Good Thinking*. Minneapolis, MN: University of Minnesota Press; 1983.
11. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Stat*. 1984;12:1151–1172.

12. Hill JR. A general framework for model-based statistics. *Biometrika* 1990;77:115–126.
13. Gelman A. Induction and deduction in Bayesian data analysis. *Rational Markets Morals*. 2011;2:67–78.
14. Leamer EE. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York, NY: John Wiley & Sons; 1978.
15. Specification Searches. Available at: http://www.anderson.ucla.edu/faculty/edward.leamer/books/specification_searches/specification_searches.htm. Accessed 8 November 2012.
16. Cornfield J. Principles of research. *Am J Ment Defic*. 1959;64:240–252.
17. Keynes JM. *A Tract on Monetary Reform*. Buffalo, NY: Prometheus Books; 2000, Ch. 3 (originally published 1923).
18. Lindley DV. A statistical paradox. *Biometrika.* 1957;44:187–192 (comment in Biometrika 1958;45: 533–534).
19. Greenland S. Multiple-bias modeling for observational studies (with Discussion). *J R Stat Soc Ser A Stat Soc*. 2005;168:267–308.
20. Bem DJ. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J Pers Soc Psychol*. 2011;100:407–425.
21. Armitage P. Comment on the paper by Lindley. *The Statistician*. 2000;51:319–320.
22. Efron B. Bayesians, frequentists, and scientists. *J Am Stat Assoc*. 2005;100:1–5.
23. Dunes DD, ed. *The Selected Writings of Benjamin Rush*. New York, NY: Rampage Press; 2007;404–418.
24. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990;263:1385–1389.
25. Phillips CV. Publication bias in situ. *BMC Med Res Methodol*. 2004;4:20.
26. Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott-Wolters-Kluwer; 2008:148–167.
27. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. *Jurimetrics*. 2011;51:113–129.
28. Greenland S. Null misinterpretation in statistical testing and its impact on health risk assessment. *Prev Med*. 2011;53:225–228.
29. Poole C. Low *P*-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291–294.
30. Gelman A, Stern HS. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat*. 2006;60:328–331.
31. Gelman A, Weakliem D. Of beauty, sex, and power: statistical challenges in estimating small effects. *Am Sci*. 2009;97:310–316.
32. Cox DR. Statistical significance tests. *Br J Clin Pharmacol*. 1982;14: 325–331.
33. Rosenthal R, Gaito J. The interpretation of levels of significance by psychological researchers. *J Psychol*. 1963;55:33–38.
34. Holman CD, Arnold-Reed DE, de Klerk N, McComb C, English DR. A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. *Epidemiology*. 2001;12:246–255.
35. Topol EJ, Moliterno DJ, Herrmann HC, et al. TARGET Investigators. Do Tirofiban and ReoPro Give Similar Efficacy Trial. Comparison of two platelet glycoprotein IIb/IIIa inhibitors, tirofiban and abciximab, for the prevention of ischemic events with percutaneous coronary revascularization. *N Engl J Med*. 2001;344:1888–1894.
36. Rabe KF. Treating COPD–the TORCH trial, *P* values, and the Dodo. *N Engl J Med*. 2007;356:851–854.
37. Crowther CA, Haslam RR, Hiller JE, Doyle LW, Robinson JS; Australasian Collaborative Trial of Repeat Doses of Steroids (ACTORDS) Study Group. Neonatal respiratory distress syndrome after repeat exposure to antenatal corticosteroids: a randomised controlled trial. *Lancet*. 2006;367:1913–1919.
38. Campion EW. Power lines, cancer, and fear. *N Engl J Med*. 1997;337: 44–46.
39. Kerlikowske K, Hubbard RA, Miglioretti DL, et al. Breast Cancer Surveillance Consortium. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: a cohort study. *Ann Intern Med*. 2011;155:493–502.
40. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet*. 2009;373:1926–1928.
41. Von Korff M, Katon WJ, Lin EH, et al. Functional outcomes of multicondition collaborative care and successful ageing: results of randomised trial. *BMJ*. 2011;343:d6612.
42. Neyman J. Frequentist probability and frequentist statistics. *Synthese*. 1977;36:97–131.
43. Fisher RA. *Statistical Methods and Scientific Inference*. 3rd ed. New York, NY: Hafner; 1973:42–43.
44. Cox DR. The role of significance tests (with discussion). *Scand J Stat*. 1977;4:49–70.
45. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ*. 2001;322:226–231.
46. Weinberg CR. It's time to rehabilitate the P-value. *Epidemiology*. 2001;12:288–290.
47. Fisher RA. Note on Dr. Berkson's criticism of tests of significance. *J Am Statist Assoc*. 1943;38:103–104. Reprinted in *Int J Epidemiol*. 2003;32:692.
48. Fisher RA. *Statistical Methods for Research Workers*. 14th ed. New York, NY: Oxford; 1970:80.
49. Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health*. 1988;78:1568–1574.
50. Royall R. *Statistical Inference: A Likelihood Paradigm*. New York, NY: Chapman and Hall; 1997.
51. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001;55:19–24.
52. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol*. 2012;22:364–368.
53. Oleckno WA. *Essential Epidemiology: Principles and Applications*. Prospect Heights, IL: Waveland Press, Inc; 2002:154.
54. Marciante KD, Bis JC, Rieder MJ, et al. Renin-angiotensin system haplotypes and the risk of myocardial infarction and stroke in pharmacologically treated hypertensive patients. *Am J Epidemiol*. 2007;166:19–27.
55. Harris M, Taylor G. *Medical Statistics Made Easy*. 2nd ed. London, UK: Scion Publishing Ltd; 2008:24–25.
56. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. 2nd ed. New York, NY: Chapman and Hall/CRC; 2003.
57. Cox DR. Another comment on the role of statistical methods. *BMJ*. 2001;322:231.
58. Shermer M. Can you hear me now?. *Scientific American*. October 2010. Available at: http://www.michaelshermer.com/2010/10/can–you–hear–me–now/. Accessed 8 November 2012.
59. Greenland S. Causal inference as a prediction problem: assumptions, identification, and evidence synthesis. In: Berzuini C, Dawid AP, Bernardinelli L, eds. *Causal Inference: Statistical Perspectives and Applications*. New York, NY: Wiley; 2012:43–58.
60. Kelly KT. Simplicity, truth, and probability. In: Bandyopadhyay,PS, Forster MR, eds. *Handbook of the Philosophy of Statistics*. North Holland, The Netherlands: Elsevier; 2011.
61. Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. *Psychol Rev*. 1963;70:193–242.
62. Cohen J. Things I have learned (so far). *Am Psychol*. 1990;45:1304–1312.
63. Loftus GR. On the tyranny of hypothesis testing in the social sciences. *Contemp Psychol*. 1991;36:102–105.
64. Hunter JE. Needed: a ban on the significance test. *Psychol Sci*. 1997; 8:3–7.
65. Jones LV, Tukey JW. A sensible formulation of the significance test. *Psychol Methods*. 2000;5:411–414.
66. Lindley DV. Lindley's paradox. *J Am Stat Assoc*. 1982;77:334–336.
67. Greenland S. Weaknesses of Bayesian model averaging for meta-analysis in the study of vitamin E and mortality. *Clin Trials*. 2009;6:42–46.
68. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY:blisher-nameSpringer; 2009.
69. Van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer; 2011.
70. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J Am Stat Assoc*. 1987;82:106–111.
71. Casella G, Berger RL. Comment. *Stat Sci*. 1987;2:344−417.
72. Altman DG. Why we need confidence intervals. *World J Surg*. 2005;29:554–556.
73. Fisher LD, van Belle G. *Biostatistics: A Methodology for the Health Sciences*. New York, NY: John Wiley & Sons; 1993:105.
74. Gerstman BB. *Epidemiology Kept Simple: An Introduction to Classic and Modern Epidemiology*. New York, NY: Wiley-Liss; 1998:169.
75. Glantz SA. *Primer of Biostatistics*. 5th ed. New York, NY: McGraw-Hill; 2001:224.

