Bayesian Methods for Evidence Evaluation: Are We There Yet?

Steven N. Goodman

Circulation. 2013;127:2367-2369; originally published online May 14, 2013;
doi: 10.1161/CIRCULATIONAHA.113.003193
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0009-7322. Online ISSN: 1524-4539

Editorial

Bayesian Methods for Evidence Evaluation
Are We There Yet?

Steven N. Goodman, MD, MHS, PhD

"First they ignore you, then they laugh at you, then they fight you, then you win," a saying reportedly misattributed to Mahatma Ghandi,[1] might apply to the use of Bayesian statistics in medical research. The idea that Bayesian approaches might be used to "affirm" findings derived from conventional methods, and thereby be regarded as more authoritative, is a dramatic turnabout from an era not very long ago when those embracing Bayesian ideas were considered barbarians at the gate. I remember my own initiation into the Bayesian fold, reading with a mixture of astonishment and subversive pleasure one of George Diamond's early pieces taking aim at conventional interpretations of large cardiovascular trials of the early 1980s.[2] It is gratifying to see that the Bayesian approach, which saw negligible application in biomedical research in the 1980s and began to get traction in the 1990s, is now not just a respectable alternative to standard methods, but sometimes might be regarded as preferable.

That said, it is premature to declare a win, and the statistical lingua franca of biomedical research is still firmly frequentist, with *P*-values, confidence intervals, and type I and II errors dominating the journal landscape. It is helpful to use the thoughtful and thorough Bayesian exercise of Bittl et al[3] to reflect on what Bayesian approaches give us, and what they do not.

Many introductions to this approach can be found in the literature, and those basics will not be repeated here.[4–6] But it is useful to examine the philosophical foundations of Bayesianism, which have their roots not so much in the Bayes theorem, from which the approach gets its name, but in its definition of probability, which Bittl et al allude to. Uncertainty can be roughly divided into 2 types: stochastic and epistemic.[7,8] Stochastic probability concerns repeatable random processes and numbers like the chance of flipping 5 successive heads. These probabilities have an objective reality (or we imagine they do), can be calculated with equations, and typically can be confirmed in simulations or empirical experiments. Epistemic probability, on the other hand, from the Greek episteme for knowledge, holds that uncertainty is a reflection of imperfect knowledge and is reflected in degree-of-belief probability statements. This uncertainty can exist about any proposition, like the chances of a defendant being guilty, the chance the United States will go to war, or, as in the article by Bittl, whether the results of randomized controlled trials (RCTs) conducted 30 years ago still apply today.

As most clearly articulated by von Mises in 1928,[9] frequentist philosophy firmly rejects such uncertainty as being representable with probabilities or, indeed, as being within the realm of science. So it is this definition of probability that we must grapple with more so than the mechanics of Bayesianism; is it scientifically legitimate to put numbers on qualitative judgments like the reliability of studies from long ago, or the comparability of studies to be pooled, the plausibility of the underlying mechanisms, or the relative reliability of RCTs versus observational studies, to conclude that percutaneous coronary intervention (PCI) is better than medical therapy and comparable to coronary artery bypass grafting (CABG) in selected patients?

It would be easy to answer no to that question if one did not have to make decisions. But patients with unprotected left main coronary artery disease must decide, together with their physicians, whether to undergo PCI or CABG. Any decision will imply a specific quantitative weighting of the evidence, or at least range of weights, even if those weights are not explicitly assigned a priori. So one cannot escape such weighting; it is either tacit, implied by the decision, or chosen a priori to reflect reasonable judgments, with the decision then following. The latter is what Bittl et al do in their Bayesian meta-analysis of the evidence supporting the relative benefits of treatments for unprotected left main coronary artery disease.[3]

Although the notion of epistemic probability might seem far afield from practice guidelines, in fact, such guidelines have their own language of epistemic uncertainty. Instead of stating that the proposition that PCI is comparable to CABG, has, say, an 80% chance of being true, the level of evidence for this statement is classified as B. The reason such level of evidence schemes have evolved is in part because traditional analyses have no measure nor language for the probability that a claim is true, even though *P*-values are occasionally misinterpreted to mean this.[10] So a body of evidence rated on the degree of potential bias in the supporting studies is put into level of evidence categories A, B, or C. The use of letters and not numbers for such distinctions is reflective of an analytic method that does not allow probabilities to be used for this purpose.

The analysis of Bittl et al differed from the standard one that motivated it in 3 ways. First was the use of Bayesian technology, next was the dimension of cross-design synthesis, and finally was the network analysis. We will discuss them in

The article about which this editorial is written appeared in the June 4, 2013 issue (*Circulation*. 2013;127:2177-2185.).

From Stanford University, Stanford, CA.

Correspondence to Steven N. Goodman, MD, MHS, PhD, Stanford University School of Medicine, Medicine, and Health Research and Policy, 259 Campus Dr, HRP/Redwood Building T265, Stanford, CA 94305-1026. Email steve.goodman@stanford.edu

(*Circulation*. 2013;127:2367-2369.)

***Circulation* is available at http://circ.ahajournals.org**
DOI: 10.1161/CIRCULATIONAHA.113.003193

that order. There have now been fairly substantive expositions in the clinical literature of what Bayesian approaches do. Bayesian analyses have at least 2 components that differ from frequentist approaches; a prior probability distribution reflecting the plausible ranges of the true effect based on external evidence, and a different calculus for how the variability in the summary estimate is calculated. This prior distribution is often chosen in a way that allows the data to speak for itself, a so-called uninformative or diffuse prior, which was used here. When there is little variation between studies, there is typically little difference between the standard random-effects model and a Bayesian summary. That is seen here, with the frequentist 95% confidence interval for the relative risk of 0.72 to 1.40 for 1-year CABG versus PCI mortality comparisons, and the Bayesian credible interval of 0.67 to 1.43, with main effect estimates near 1. In general, the variability of a Bayesian meta-analytic estimate will be wider than with a frequentist analysis, and will more accurately measure an estimate's true uncertainty. In this example, if we look solely at the numbers provided, the Bayesian approach appears to yield little different than the standard method.

The cross-design synthesis is next. Although the name sounds technical, it really is nothing more than pooling the results of studies with different designs, as is done here with the RCTs and 2 kinds of cohort studies. If the studies are estimating exactly, the same quantity (eg, a relative risk), there is in theory nothing wrong with this. But studies of different design typically have different susceptibility to biases, with consequently differing confidence in the accuracy of their results. How do we reflect that differing confidence? One way is to use meta-regression to examine how the study estimates vary as a function of design features, although there are typically too few studies for this to tell us enough. However this is done, the net effect is to downweight studies that we believe are more prone to bias. In the absence of adequate empirical data, we can simply choose weights that reflect informally what we believe to be the relative credibility of a given design to the most rigorous design in the synthesis, typically an RCT. The article describes this approach, but not in enough detail to reproduce, ie, with the quantitative details of the models. There is also not enough qualitative detail about the underlying studies to independently assess their bias potential, although some of that information is in the original American College of Cardiology guideline document.[11] Given that the effects of RCTs, cohort, and matched cohort studies were statistically similar, the result in this example could not have been sensitive to these weights, but the details could still be useful to inform future analyses.

Finally, we come to the network meta-analysis, only, in this case, there was not much of a network, with only 2 pairs of comparisons. With strong consistent evidence that CABG is equivalent to PCI (for mortality), and strong consistent evidence that CABG is better than medical therapy, it does not take a sophisticated analysis to confirm the conclusion that PCI is likely better than medical therapy, although the exact effect estimates and their variability will not be as apparent. The main complication was that almost all of the CABG versus medical therapy studies were decades in the past, and both therapies have improved over time. The authors deal with this in a variety of reasonable ways, allowing different variation of medical therapy versus CABG outcomes, because medical therapy has probably improved more over time, and reducing the weight of the older studies by a factor of 3, which they varied in sensitivity analyses.

It is here, in particular, that it would have been nice to have seen the authors take more advantage of the Bayesian technology and measures. Stopping with effect estimates and credible intervals essentially shackles their Bayesian analysis with some of the same limitations of a frequentist report. It would have been very interesting to have calculated the probability of equivalence (within defined limits) between CABG and PCI, of the difference between PCI and medical therapy, and perhaps Bayes factors for such quantities.[4,12] Having implemented approaches to downweight for study age and risk of bias, it would have been quite interesting to see how these affected estimates of the probability of effectiveness. Diamond and Kaul[13] have outlined an alternative approach that would replace level of evidence designations with a completely Bayesian calculus that classifies conclusions instead into the categories used in legal settings, namely, beyond a reasonable doubt, clear and convincing evidence, etc. Although their specific approach raised other issues,[14] it shows how the probabilistic conclusions of a Bayesian analysis can link with the levels of evidence used in clinical guideline development. That final step would have been welcome here, highlighting what Bayesian approaches have to offer in support of guidelines that simply cannot be matched by frequentist summaries.

Another advantage of Bayesian analyses for guideline development is transparency. Diamond and many others have decried the sometimes hidden judgments that underlie guidelines, particularly those not driven by the highest level RCT evidence. Bayesian approaches, which force the quantification of these judgments and model their consequences, can make evidential assessment behind guidelines more transparent and reproducible, allowing us to explicate and perhaps narrow the areas of disagreement. Although these authors have done this kind of modeling, the details and sensitivity analyses are not presented in enough detail to reproduce them.

Settings in which one would expect to see bigger differences between Bayesian and frequentist approaches in the numeric estimates would be those in which there was some previous information, substantive quantitative and qualitative variability among studies, and a richer web of treatment comparisons from which to derive indirect inferences. Less obvious, but no less important, is the value of such models for modeling the effects of different qualitative judgments and allowing guideline panels to navigate more adeptly in the world of intermediate certainty. This analysis by Bittl et al[3] can serve as a nice starting point for future Bayesian evidence analyses, whose value may be more apparent when the outcomes of the summaries are less predictable, qualitative disagreements among experts about the evidence are more difficult to reconcile, and the full range of Bayesian measures is used.

Disclosures

None.

References

1. O'Carroll E. Political misquotes: the 10 most famous things never actually said. *Christian Science Monitor*. June 3, 2011. http://www.csmonitor.com/USA/Politics/2011/0603/Political-misquotes-The-10-most-famous-things-never-actually-said/ Accessed on May 13, 2013.
2. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Intern Med*. 1983;98:385–394.
3. Bittl J, He Y, Jacobs A, Yancy C, Normand S-L. Bayesian methods affirm the use of percutaneous coronary intervention to improve survival in patients with unprotected left main coronary artery disease. *Circulation*. 2013;127:2177–2185.
4. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130:1005–1013.
5. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA*. 1995;273:871–875.
6. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006;35:765–775.
7. Oakes M. *Statistical Inference: A Commentary for the Social Sciences*. New York: Wiley; 1986.
8. Goodman SN. Probability at the bedside: the knowing of chances or the chances of knowing? *Ann Intern Med*. 1999;130:604–606.
9. von Mises R. *Probability, Statistics and Truth*. 2nd rev. English ed. New York: Dover; 1981.
10. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008;45:135–140.
11. Levine GN, Bates ER, Blankenship JC, Bailey SR, Bittl JA, Cercek B, Chambers CE, Ellis SG, Guyton RA, Hollenberg SM, Khot UN, Lange RA, Mauri L, Mehran R, Moussa ID, Mukherjee D, Nallamothu BK, Ting HH. 2011 ACCF/AHA/SCAI Guideline for Percutaneous Coronary Intervention: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. *Circulation*. 2011;124:e574–e651.
12. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90:773–795.
13. Diamond GA, Kaul S. Bayesian classification of clinical practice guidelines. *Arch Intern Med*. 2009;169:1431–1435.
14. Goodman SN. Building a Bayesian bridge from evidence to guidelines. *Arch Intern Med*. 2009;169:1436–1437.

KEY WORDS: Editorials ■ Bayes Theorem ■ biostatistics ■ guideline ■ methodology