The Value of Bayesian Methods for Evidence-Based Medicine

Steven Goodman, MD, MHS, PhD
Professor of Medicine and Epidemiology
Co-director, METRICS
(Meta-research Innovation Center at Stanford)

FDA Workshop on Substantial Evidence
April 21, 2016
"Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision Making?"

5/20/2004
Norris Aldersen
Janet Woodcock
Bob Temple
Steve Goodman
Greg Campbell
Don Berry
Don Rubin
Tom Louis
Larry Kessler
Telba Irony
Constantine Gatsonis
Janet Wittes
Andy Grieve
Frank Rockhold
Bill DuMouchel
Ralph D'Agostino
Susan Ellenberg
Stacy Lindborg
Robert O'Neill
Jay Siegel

Table of Contents
August 2005; 2 (4)

Articles

Norris E Alderson
Editorial
Full Text (PDF)  Request Permissions

Janet Woodcock
FDA introductory comments: clinical studies design and evaluation issues
Full Text (PDF)  References  Request Permissions

Robert Temple
How FDA currently makes decisions on clinical studies
Full Text (PDF)  References  Request Permissions

Steven N Goodman
Introduction to Bayesian methods I: measuring the strength of evidence
The Rational Clinical Examination

Evidence-Based Medicine
A New Approach to Teaching the Practice of Medicine

Evidence-Based Medicine Working Group

A NEW paradigm for medical practice is emerging. Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the physician, including efficient literature searching and the application of formal rules of evidence evaluating the clinical literature.

dose of phenytoin intravenously and the drug is continued orally. A computed tomographic head scan is completely normal, and an electroencephalogram shows only nonspecific findings. The patient is very concerned about his risk of seizure recurrence. How might the resident proceed?

The Way of the Past

Faced with this situation as a clinical clerk, the resident was told by her co-
year is between 43% and 51%, and at 3 years the risk is between 51% and 60%. After a seizure-free period of 18 months his risk of recurrence would likely be less than 20%. She conveys this information to the patient, along with a recommendation that he take his medication, see his family doctor regularly, and have a review of his need for medication if he remains seizure-free for 18 months. The patient leaves with a clear idea of his likely prognosis.
Definition of EBM

1992: Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the physician, including efficient literature searching and the application of formal rules of evidence evaluating the clinical literature.

2000: A systematic approach to clinical problem solving which allows the integration of the best available research evidence with clinical expertise and patient values.
Which candidate compound would be best in a future population?

Percent of subjects (N=200) for which compound was favored

- Candidate Compound number 1: 28%
- Candidate Compound number 2: 14%
- Candidate Compound number 3: 9.3%
- Candidate Compound number 4: 6.8%
- Candidate Compound number 5: 6.5%
- Candidate Compound number 6: 5.5%

P = 0.003
Which candidate would be best in a future population?

Percent of respondents (N=200) for which candidate was favored:

- Trump: 28%
- Carson: 14%
- Bush: 9.3%
- Rubio: 6.8%
- Cruz: 6.5%
- Fiorina: 5.5%

Which candidate would be best in a future population?
Things identified as cancer risks
(JNCI news, 1992)

- Electric Razors
- Broken Arms
  (in women)
- Fluorescent lights
- Allergies
- Breeding Reindeer
- Being a waiter
- Owning a pet bird
- Hot dogs
- Being short
- Being tall

Having a refrigerator
Magnets Lessen Foot Pain Of Diabetics, a Study Finds

By HOLCOMB B. NOBLE

In one of the first scientific studies of the centuries-old and highly debated use of magnets for treatment of medical disorders, a New York neurologist reported today that he had significantly lessened the foot pain that afflicts millions of diabetics.

Dr. Michael I. Weintraub, a clinical professor of neurology at New York Medical College, emphasized that his study was small, involving only 24 patients, and must be regarded as preliminary to much more clinical research. But he said that the early results were clear and that the treatment ought to be put to use immediately, provided the correct magnets are used and the treatment is limited to the types of pain that have been studied.

The study, which appears in this month’s issue of the American Journal of Physical Medicine, was conducted at the Veterans Administration Medical Center in New York City in the late 1980s.

Dr. Weintraub said he applied low-intensity magnets to the feet of patients with foot pain for 20 minutes, three times a week, for six weeks.

“The early results were clear and... the treatment ought to be put to use immediately,” Dr. Weintraub said.

“We have no idea how or why the magnets work.”

“A real breakthrough…”

“...the [study] must be regarded as preliminary.”

“But...the early results were clear and... the treatment ought to be put to use immediately.”
Is everything we eat associated with cancer? A systematic cookbook review¹-³

Jonathan D Schoenfeld and John PA Ioannidis

ABSTRACT
Background: Nutritional epidemiology is a highly prolific field. Debates on associations of nutrients with disease risk are common in the literature and attract attention in public media.
Objective: We aimed to examine the conclusions, statistical significance, and reproducibility in the literature on associations between specific foods and cancer risk.
Design: We selected 50 common ingredients from random recipes in a cookbook. PubMed queries identified recent studies that evaluated the relation of each ingredient to cancer risk. Information regarding author conclusions and relevant effect estimates were extracted. When >10 articles were found, we focused on the 10 most recent articles.
Results: Forty ingredients (80%) had articles reporting on their cancer risk. Of 264 single-study assessments, 191 (72%) concluded that the tested food was associated with an increased (n = 103) or a decreased (n = 88) risk; 75% of the risk estimates had weak (0.05 > P ≥ 0.001) or no statistical (P > 0.05) significance. Statistically significant results were more likely than nonsignificant findings to be published in the study abstract than in only the full text (P < 0.0001). Meta-analyses (n = 36) presented more conservative results; only 13 (26%) reported an increased (n = 4) or a decreased (n = 9) risk (6 had more than weak statistical support). The median RRs (IQRs) for studies that concluded an increased or a decreased risk were 2.20 (1.60, 3.44) and 0.52 (0.39, 0.66), respectively. The RRs from the meta-analyses were on average null (median: 0.96; IQR: 0.85, 1.10).
Conclusions: Associations with cancer risk or benefits have been claimed for most food ingredients. Many single studies highlight implausibly large effects, even though evidence is weak. Effect sizes shrink in meta-analyses. Am J Clin Nutr 2013;97:127–34.

and such discrepancies in the evidence have fueled hot debates (9–12) rife with emotional and sensational rhetoric that can subject the general public to increased anxiety and contradictory advice (13, 14). One wonders whether this highly charged atmosphere and extensive testing of food-related associations may create a plethora of false-positive findings (15) and questionable research practices, especially when the research is highly exploratory, the analyses and protocols are not preregistered, and the findings are selectively reported. It was previously shown in a variety of other fields that “negative” results are either less likely to be published (16–21) or misleadingly interpreted (19, 22). Studies may spuriously highlight results that barely achieve statistical significance (15, 23) or report effect estimates that either are overblown (24, 25) or cannot be replicated in other studies (24, 26, 27).

To better evaluate the extent to which these factors may affect studies investigating dietary risk factors for malignancy, we surveyed recently published studies and meta-analyses that addressed the potential association between a large random sample of food ingredients and cancer risk of any type of malignancy.

SUBJECTS AND METHODS

Random ingredient selection

We selected ingredients from random recipes included in The Boston Cooking-School Cook Book (28), available online at http://archive.org/details/bostoncookingsch00farmrich. A copy of the book was obtained in portable document format and viewed by using Skim version 1.3.17 (http://skim-app.sourceforge.net). The recipes (see Supplementary Table 1 under “Supplemental data” in the online issue) were selected at random by generating a random number corresponding to each book...
**Design:** We selected 50 common ingredients from random recipes in a cookbook. PubMed queries identified recent studies that evaluated the relation of each ingredient to cancer risk. Information regarding author conclusions and relevant effect estimates were extracted. When >10 articles were found, we focused on the 10 most recent articles.

**Results:** Forty ingredients (80%) had articles reporting on their cancer risk. Of 264 single-study assessments, 191 (72%) concluded that the tested food was associated with an increased \((n = 103)\) or a decreased \((n = 88)\) risk; 75% of the risk estimates had weak (0.05
A well done study is reported on a new electrical stimulator for sickle cell pain control, and the authors state that it has turned out, somewhat surprisingly (e.g. < 20% chance of being true before the experiment), to be effective in reducing migraine frequency by 15%, 95% CI: 1% to 29%, $P=0.03$. The probability that this association is real is:

a.) < 75%

b.) 75% to 94.99...%

c.) 95% or higher
A short research quiz

A well done study is reported on a new electrical stimulator for sickle cell pain control, and the authors state that it has turned out, somewhat surprisingly (e.g. < 20% chance of being true before the experiment), to be effective in reducing migraine frequency by 15%, 95% CI: 1% to 29%, $P=0.03$. The probability that this association is real is:

- a.) < 75%
- b.) 75% to 94.99...%
- c.) 95% or higher
Physicians must keep current with clinical information to practice evidence-based medicine (EBM). In doing so, most prefer to seek evidence-based summaries, which give the clinical bottom line, or evidence-based practice guidelines. Resources that maintain these information summaries, however, currently include a limited number of common conditions. Thus, to answer many of their clinical questions, physicians need to access reports of original research. This requires the reader to critically appraise the design, conduct, and analysis of each study and subsequently interpret the results.

Several surveys in the 1980s demonstrated that practicing physicians, particularly those with no formal education in epidemiology and biostatistics, had a poor understanding of common statistical tests and limited ability to interpret study results. Many physicians likely have increased difficulty today because more complicated sta-
277 Yale Residents
95% felt it was important to understand these concepts to be an intelligent reader of the literature.

Average score: 41%
Irony...

- 59% reportedly chose the correct “P-value” definition, but all of the multiple choice answers were incorrect!
Statisticians issue warning over misuse of $P$ values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

07 March 2016

Misuse of the $P$ value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warns in a statement released today\(^1\). The group has taken the unusual step of issuing principles to guide use of the $P$ value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society’s members had become increasingly concerned that the $P$ value was being misapplied in ways that cast doubt on statistics generally, he adds.

In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions based on $P$ values alone. Researchers should describe not only the data analyses that produced statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely
Statisticians Found One Thing They Can Agree On: It’s Time To Stop Misusing P-Values

By CHRISTIE ASCHWANDEN

Little p-value
What are you trying to say
Of significance?
Maintenance Therapy With Tumor-Treating Fields Plus Temozolomide vs Temozolomide Alone for Glioblastoma: A Randomized Clinical Trial

Roger Stupp, MD; Sophie Taillibert, MD; Andrew A. Kanner, MD; Santosh Kesari, MD, PhD; David M. Steinberg, PhD; Steven A. Toms, MD, FACS, MPH; Lynne P. Taylor, MD, FAAN; Frank Lieberman, MD; Antonio Silvani, MD; Karen L. Fink, MD, PhD; Gene H. Barnett, MD, MBA; Jay-Jiguang Zhu, MD, PhD; John W. Henson, MD, MBA, FAAN; Herbert H. Engelhard, MD, PhD; Thomas C. Chen, MD, MSc; Joseph Landolfi, DSc; Eilon D. Kirson, MD

MAIN OUTCOMES AND MEASURES The primary end point was progression-free survival in the intent-to-treat population (significance threshold of .01) with overall survival in the per-protocol population (n = 280) as a powered secondary end point (significance threshold of .006). This prespecified interim analysis was to be conducted on the first 315 patients after at least 18 months of follow-up.

RESULTS The interim analysis included 210 patients randomized to TTFields plus temozolomide and 105 randomized to temozolomide alone, and was conducted at a median follow-up of 38 months (range, 18-60 months). Median progression-free survival in the intent-to-treat population was 7.1 months (95% CI, 5.9-8.2 months) in the TTFields plus temozolomide group and 4.0 months (95% CI, 3.3-5.2 months) in the temozolomide alone group (hazard ratio [HR], 0.62 [98.7% CI, 0.43-0.89]; P = .001). Median overall survival in the per-protocol population was 20.5 months (95% CI, 16.7-25.0 months) in the TTFields plus temozolomide group (n = 196) and 15.6 months (95% CI, 13.3-19.1 months) in the temozolomide alone group (n = 84) (HR, 0.64 [99.4% CI, 0.42-0.98]; P = .004).

CONCLUSIONS AND RELEVANCE In this interim analysis of 315 patients with glioblastoma who had completed standard chemoradiation therapy, adding TTFields to maintenance temozolomide chemotherapy significantly prolonged progression-free and overall survival.
Figure 2. Survival Curves for Patients Included in the Interim Analysis in the Intent-to-Treat Population

A. Progression-free survival

HR, 0.62 (98.7% CI, 0.43-0.89); log-rank P = .001

TTFields plus temozolomide

Temozolomide alone

B. Overall survival

HR, 0.74 (95% CI, 0.56-0.98); log-rank P = .03

TTFields plus temozolomide

Temozolomide alone

No. at risk

TTFields plus temozolomide: 210, 149, 94, 60, 45, 35, 29, 22, 16, 12, 11

Temozolomide alone: 105, 55, 26, 21, 15, 12, 6, 5, 1, 1

Survival analyses on time from date of randomization until tumor progression, death, or last follow-up (censored patients) according to the Kaplan-Meier method. The small vertical ticks on the curves indicate censored patients. HR indicates hazard ratio; TTFields, tumor-treating fields.
Rationale: Prior evidence

Tumor-treating fields (TTFIELDS) are an antimitotic treatment that selectively disrupts the division of cells by delivering low-intensity, intermediate-frequency (200 kHz) alternating electric fields via transducer arrays applied to the shaved scalp.\textsuperscript{8-10} In preclinical models, TTFIELDS have been shown to cause mitotic arrest and apoptosis by disrupting mitotic spindle formation during metaphase and causing dielectrophoretic movement of polar molecules during cytokinesis.\textsuperscript{8,10-12} In a randomized phase 3 trial in which TTFIELDS were compared with chemotherapy in 237 patients with recurrent glioblastoma, the use of TTFIELDS did not prolong progression-free survival or overall survival, but the therapy was associated with better quality of life without the typical chemotherapy-associated toxic effects.\textsuperscript{13}
The propagation of the strength of evidence is through a Bayesian algorithm, with the strength of evidence represented by the degree to which the probability of a clinical effect is modified by evidence from the component steps. This modeling makes clear how strong mechanistic evidence can be necessary for proper inferences, yet still, by itself, yield very low probabilities of success for a given intervention.
Again, the basic research process has increased the odds of success more than does the clinical research process, ….. So while the translational and developmental process is often decried for its small yield of usable therapies, it must be recognized that it increases the odds of success by several orders of magnitude, leaving the clinical evaluation process to increase it yet further to justify clinical use.

The propagation of the strength of evidence is through a Bayesian algorithm, with the strength of evidence represented by the degree to which the probability of a clinical effect is modified by evidence from the component steps. This modeling makes clear how strong mechanistic evidence can be necessary for proper inferences, yet still, by itself, yield very low probabilities of success for a given intervention.
## P-values: Bayesian Translations

<table>
<thead>
<tr>
<th>P-value (Z-score)</th>
<th>Minimum Bayes factor</th>
<th>(-e \cdot p \ln(p))</th>
<th>Strength of evidence</th>
<th>Decrease in probability of the null hypothesis, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.10) (1.64)</td>
<td>.26</td>
<td>.6</td>
<td>Weak</td>
<td>From 75, To no less than 44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 50, To 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 17, To 5</td>
</tr>
<tr>
<td>(0.05) (1.96)</td>
<td>.15</td>
<td>.4</td>
<td>Weak to Moderate</td>
<td>From 75, To no less than 31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 50, To 13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 26, To 5</td>
</tr>
<tr>
<td>(0.03) (2.17)</td>
<td>.1</td>
<td>.3</td>
<td>Moderate</td>
<td>From 75, To no less than 22 - 47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 50, To 9 - 23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 33, To 5 - 13</td>
</tr>
<tr>
<td>(0.01) (2.58)</td>
<td>.04</td>
<td>.13</td>
<td>Moderate to strong</td>
<td>From 75, To no less than 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 50, To 3.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 60, To 5</td>
</tr>
<tr>
<td>(0.001) (3.28)</td>
<td>.005</td>
<td>.02</td>
<td>Strong to very strong</td>
<td>From 75, To no less than 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 50, To 0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>From 92, To 5</td>
</tr>
</tbody>
</table>
$P = 0.05$

(Strongest BF = 0.14)

Probability of truth

- 1% -> 6%
- 10% -> 43%
- 25% -> 69%
- 50% -> 87%
- 100%
P=0.01  
(strongest BF =0.04)  

Probability of truth
Scientific claims are not...

Rather...

0% to 100% uncertain
Borrowing information from adults to children: A Case Study of Guillan-Barré Syndrome
Adult vs. Pediatric GBS

- No known difference in pathophysiology or clinical course, except that children recover more quickly, almost never die, and have fewer serious sequelae.
What do we know in adults?: PE vs. Placebo

- Two RCTs of plasma exchange versus placebo showed identical effects on median time to unaided walking (Time to grade 2):

<table>
<thead>
<tr>
<th>Study</th>
<th>N (age)</th>
<th>Placebo</th>
<th>PE</th>
<th>HR</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>French GBS Study (1985)</td>
<td>245 (&gt;16)</td>
<td>111 d</td>
<td>70 d</td>
<td>0.63</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GBS Study Group (1985)</td>
<td>220 (&gt;12)</td>
<td>85 d</td>
<td>53 d</td>
<td>0.62</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
### What do we know in children? PE vs. Historical Control Studies

<table>
<thead>
<tr>
<th>Author</th>
<th>Mean treated (N)</th>
<th>Mean untreated (N)</th>
<th>HR</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epstein 1990</td>
<td>24 (9)</td>
<td>60 (14)</td>
<td>0.4</td>
<td>0.17-0.93</td>
</tr>
<tr>
<td>Lamont 1991</td>
<td>17 (6)</td>
<td>43 (18)</td>
<td>0.4</td>
<td>0.15 - 0.99</td>
</tr>
<tr>
<td>Jansen 1993</td>
<td>16 (8)</td>
<td>29 (11)</td>
<td>0.55</td>
<td>0.23 - 1.37</td>
</tr>
<tr>
<td>Graf 1999</td>
<td>76 (6)</td>
<td>50 (9)</td>
<td>1.52</td>
<td>0.54 - 4.3</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>RE model</strong></td>
<td><strong>0.58</strong></td>
<td><strong>0.32-1.0</strong></td>
</tr>
</tbody>
</table>
What do we know in Adults?:
PE vs. IV Ig

PSGBS Study Group,
Lancet 1997; 349: 225-230

Median time to walking:
IVIg: 51 d
PE: 49 d
IVIg+PE: 45 d

HR IVIG vs PE:
1.04, CI 0.8 to 1.4
What do we know in Adults?:
PE vs. IV Ig

Van der Meché, NEJM 1992

Median time to walking:
IVIg: 55 d (N=74)
PE: 69 d (N=73)

HR: 0.80, CI 0.62 to 1.02, p=0.07

Meta-analytic average of two PE vs. IVIg studies:
HR = 0.91, CI 0.75 to 1.1
Prior on IVIG vs PE Effect

Hazard ratio
(ratio of median recovery times)
The value of borrowing

- Save 80% of the sample size of a new trial.
- Define the amount of evidence needed to be convincing (BF=3), on top of the adult evidence.
- Promote conversations among scientists about the things they know something about (e.g. similarity of clinical course and pathophysiology)
L. Moyé, MD, PhD

“What we have to wrestle with is how to interpret p-values for secondary endpoints in a trial which frankly was negative for the primary. ...In a trial with a positive endpoint...you haven’t spent all of the alpha on that primary endpoint, and so you have some alpha to spend on secondary endpoints....In a trial with a negative finding for the primary endpoint, you have no more alpha to spend for the secondary endpoints.”
Dr. Lipicky: What are the p-values needed for the secondary endpoints? ...Certainly we’re not talking 0.05 anymore. ...You’re out of this 0.05 stuff and I would have like to have seen what you thought was significant and at what level...

What p-value tells you that it’s there study after study?

Dr. Konstam: …what kind of statistical correction would you have to do that survival data given the fact that it’s not a specified endpoint? I have no idea how to do that from a mathematical viewpoint.
The sine qua non of EBM is integration of evidence of different types; empirical, mechanistic, clinical.

Without Bayesian formalism and measures, there is no quantitative language, no conceptual framework, and barely even a qualitative language to do this.

The social-scientific conventions in current use (e.g. evidence hierarchies and $P \leq 0.05$, Power $> 80\%$) are unmoored from either proper measures of evidential strength or probability of truth. These conventions have also inhibited progress in developing new standards.

Meta-research can be terrifically informative regarding the reliability of scientific methods.

The needs and role of FDA are not the same as an EBM practitioner, but it can play a huge role in helping to establish new evidential standards and a regulatory framework for the 21st century that reflects a coherent and consistent approach to inference.
Lectures Aren't Just Boring, They're Ineffective, Too, Study Finds

12 May 2014 3:00 pm | 122 Comments

Blah? Traditional lecture classes have higher undergraduate failure rates than those using active learning techniques, new research finds.

Are your lectures droning on? Change it up every 10 minutes with more active teaching techniques and more students will succeed, researchers say. A new study finds that
Some studies aren't worth stressing over.

Scientific studies about how people act or think can rarely be replicated by outside experts, said a study Thursday that raised new questions about the seriousness of psychology research.

A team of 270 scientists tried reproducing 100 psychology and social science studies that had been published in three top peer-reviewed U.S. journals in
Thank you.