

Different Strokes for Different Folks

A Bayesian Approach to Heterogeneous Treatment Effects in Empirical Cost-Effectiveness Analysis

Dave Vanness, Ph.D

AcademyHealth
Annual Research Meeting 2007



“Now, the world don't move to the beat of just one drum.

What might be right for you, may not be right for some.”

J Clin Epidemiol. Vol. 49, No. 4, pp. 395-400, 1996
Copyright © 1996 Elsevier Science Inc.



S0895-4166(96)0115-0
S0895-4166(96)0115-0

Can Treatment That Is Helpful on Average Be Harmful to Some Patients? A Study of the Conflicting Information Needs of Clinical Inquiry and Drug Regulation

Ralph J. Horwitz,^{1,2} Burton H. Singer,³
Robert W. Makuch,² and Catherine M. Viscoli¹

JGIM September 15, 2006—Vol 21, No 10

©2006 American Medical Association. All rights reserved.

COMMENTARY

Moderators of Treatment Outcomes
Clinical, Research, and Policy Importance

Helena C. Kramer, PhD

Ellen Frank, PhD

David J. Kopelman, MD

JGIM September 15, 2006—Vol 21, No 10

Moderators of Treatment Outcomes
Clinical, Research, and Policy Importance

Helena C. Kramer, PhD
Ellen Frank, PhD
David J. Kopelman, MD

What any patient wants to know is “Would this treatment elicit a better response than the control for *me in particular?*”

The American Journal of Medicine (2007) Vol 120 (4A), 51-59



THE AMERICAN
JOURNAL of
MEDICINE

Heterogeneity of Treatment Effects: Implications for Guidelines, Payment, and Quality Assessment

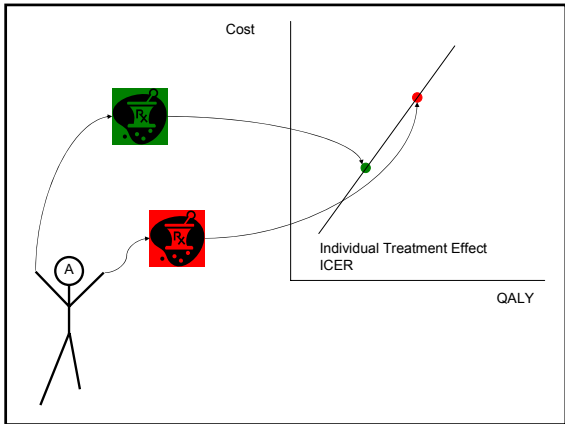
Sheldon Greenfield, MD,* Richard Kravitz, MD, MSPH,* Nathua Duan, PhD,* and Sherrie H. Kaplan, PhD, MPH*

© 2009 Blackwell Publishing Ltd, 978-1-4447-0132-1, 1-10
THE ASSOCIATION OF MEDICAL PROFESSIONS
Heterogeneity of Treatment Effects: Implications for Guidelines, Payment, and Quality Assessment
Jonathan Doolittle, MD, MPH; Richard Goepfert, MD, MPH; Andrew Auer, PhD; and Steven D. Kaplan, PhD, MPH

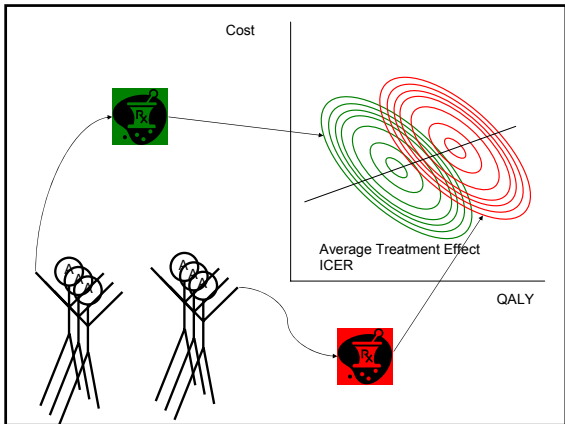
Even more disturbing than the implications of HTE for clinical practice is the averaging of average effects across trials, as is done in the current evidence-based medicine approach to quality improvement, including the development of clinical guidelines, quality-of-care measures, and pay-for-performance initiatives. These initiatives promote a sweeping uniformity of treatment that will almost certainly cause eventual harm.

Basic Terminology ...
 with stick-figure art

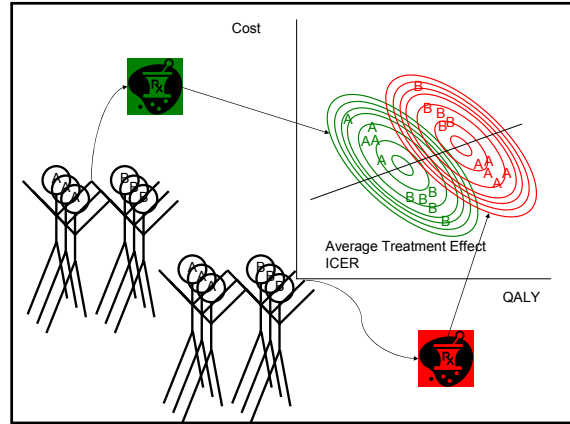
Pure Counterfactual Trial



Idiosyncratic Heterogeneity
 (Chance)



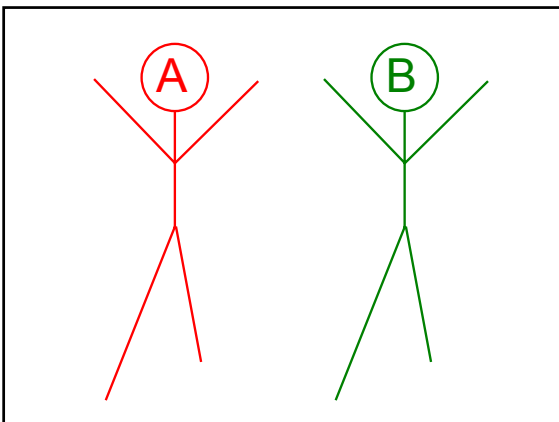
Essential Heterogeneity (Latent Class)



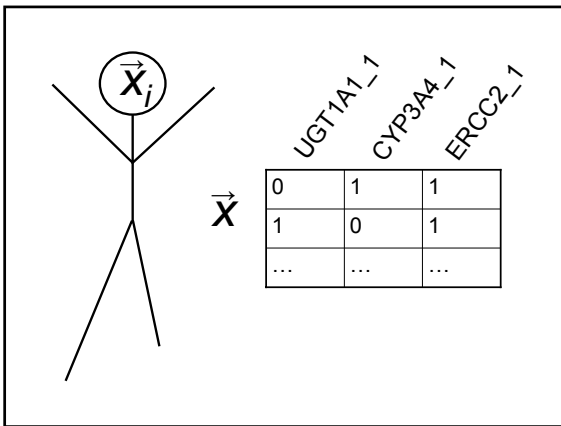
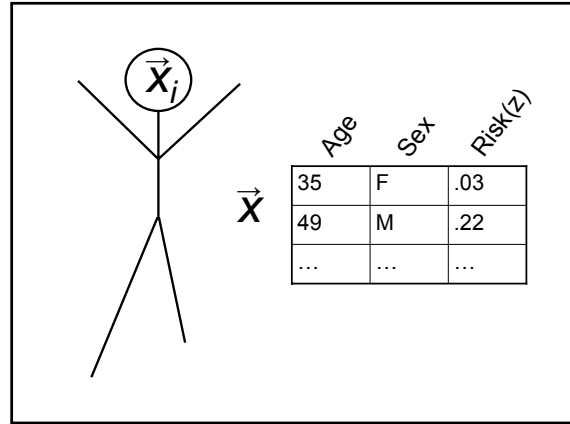
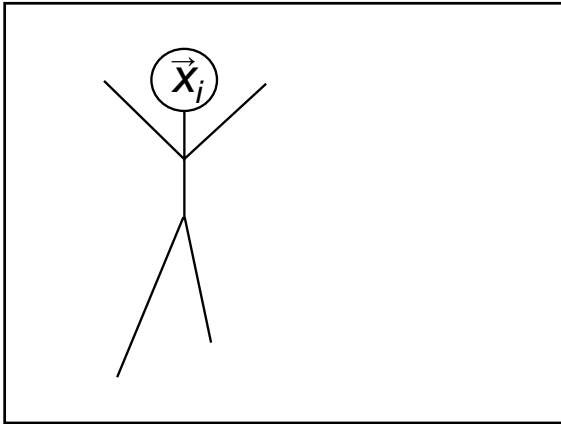
Moderators of Treatment Outcomes
Clinical, Research, and Policy Importance

A common, but generally counterproductive, approach is to control for baseline factors. Interactions are typically excluded from such analyses, imposing the assumption that these are not treatment moderators. The effect size before adjustment is the effect size in the total population. If population interactions are absent, the effect size after adjustment estimates the common effect size in subgroups matched on those factors. However, if there is no common effect size in the population, the effect size after adjustment is generally uninterpretable.

If we could observe class...



But what if we only observe...



Pharmacogenomics

- Individual genetic variations in metabolism, transport or cellular target of a drug affects toxicity or therapeutic effect (McLeod and Evans, 2001)
- High throughput techniques for identifying variations are now becoming widely available (Marsh et al, 2002) offering potential for individually-tailored therapies.
- While some individual single nucleotide polymorphisms (SNPs) have been associated with heterogeneous treatment effects, pharmacogenomics has not yet revolutionized the practice of medicine.

Approaches to Classification

- There are many approaches to identifying latent class membership based on interactions of observed variables (for a nice, free introduction, see Magidson and Vermunt, available online at <http://www.statisticalinnovations.com/articles/sage11.pdf>)

Recursive partitioning is a non-parametric method that repeatedly splits populations into groups with maximally dissimilar outcomes of interest (see Zhang and Singer, Recursive Partitioning in the Health Sciences, Springer 2005), forming "trees" of interacted variables.

Tradeoffs exist between goodness of fit and tree complexity (number of splits).

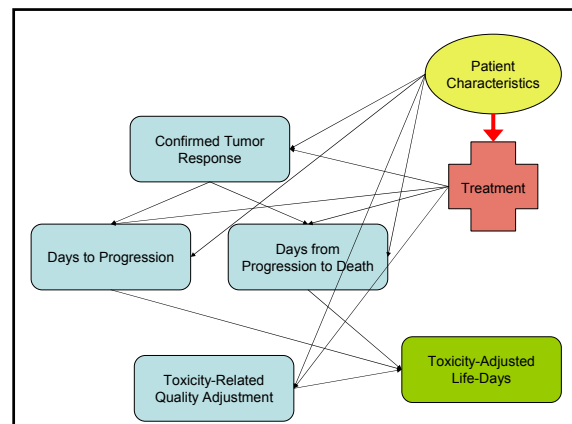
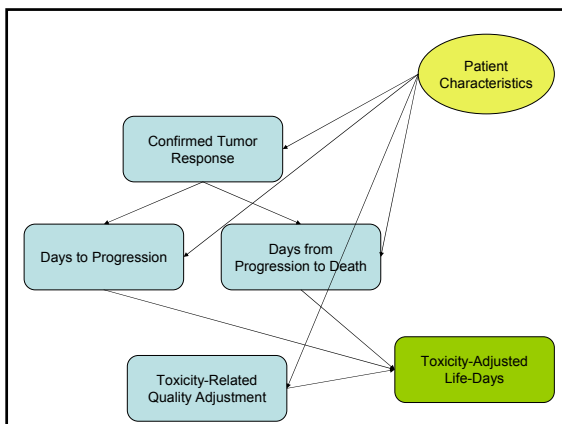
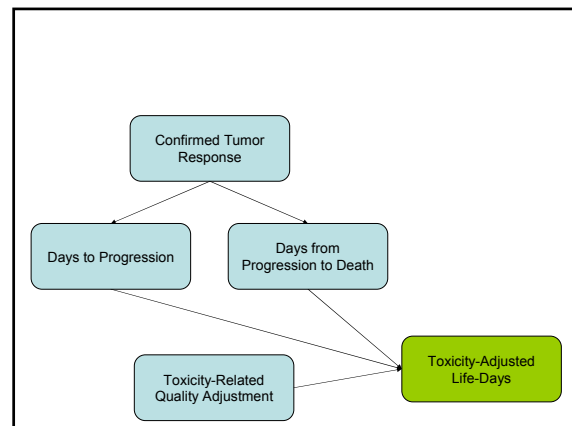
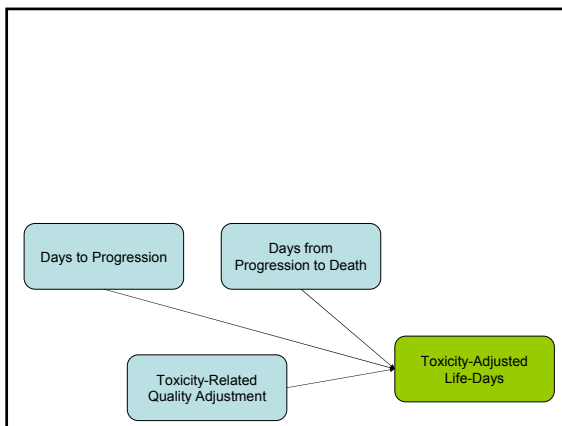
A real-world example using recursive partitioning to identify potential heterogeneous treatment effects...

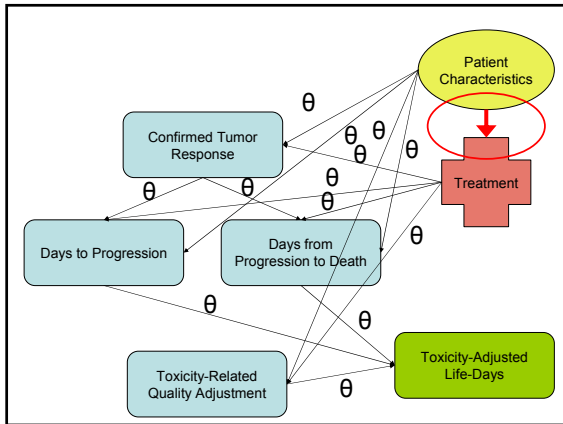
Work in progress: please do not cite without permission...

Data

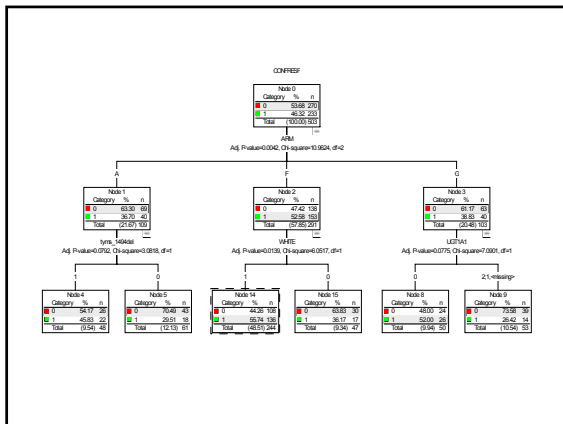
- Data comes from 503 participants from a clinical trial of adjuvant chemotherapy for advanced colorectal cancer who provided blood samples for genomic screening
- Patients were randomly assigned to three treatment regimens (A, F, G)
- Dates of progression and/or death were imputed for those who were alive and/or non-progressed at last follow-up

Toxicity-Adjusted
Life-Days





Proposed Heterogeneous Treatment Effect Groups for Confirmed Tumor Response



Bayesian Posterior Distributions of Unknown Parameters (Heterogeneous Treatment Effects on Confirmed Response) Conditional on Observed Data, estimated with WinBUGS 1.4.1

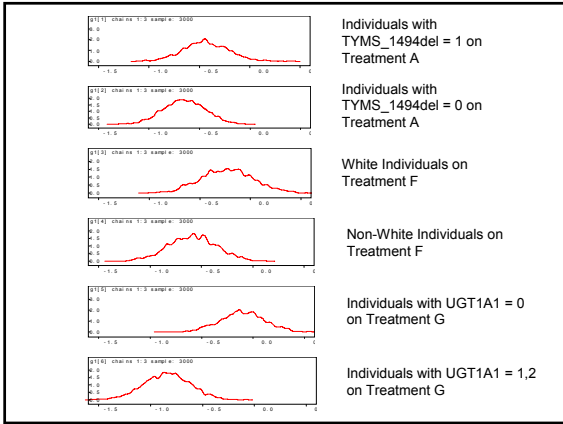
Bayesian Posterior Distributions of Unknown Parameters (Heterogeneous Treatment Effects on Confirmed Response) Conditional on Observed Data, estimated with WinBUGS 1.4.1

FREE!!!

<http://www.mrc-bsu.cam.ac.uk/bugs/>

Bayes Rule

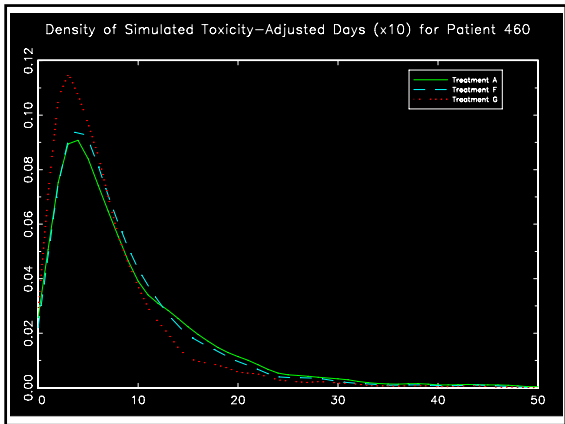
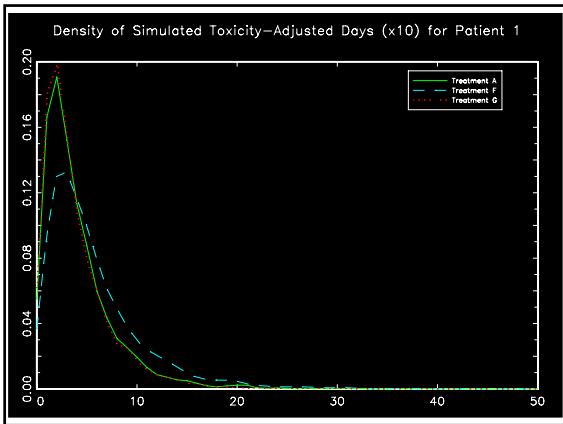
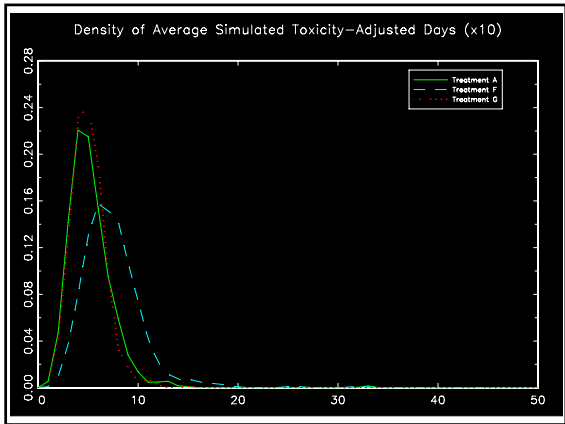
- $P(\theta|Data) = L(Data|\theta) P(\theta) + P(X)$
- $P(\theta|Data)$ is called the "posterior" – it summarizes our uncertainty about θ (the true relationship between X and outcomes) after observing the data.
- $P(\theta)$ is our "prior" knowledge about θ (it can come from meta-analysis of previous studies, or can be made "non-informative" (flat), or can accommodate elicited expert opinion).
- $L(Data|\theta)$ is the same likelihood function we use in classical statistics. It represents the model for how the data were generated given the true relationship between X and outcomes.
- In Bayesian Analysis, the object of inference is θ , not Data



Simulated Treatment

- Take a draw of all unknown θ from the posterior
 - represents our uncertainty about relationship among covariates, treatments and outcomes
- For each individual in the dataset (or for a hypothetical individual)
 - Take their actual covariate values (age, sex, polymorphism profile)
 - “Assign” treatment by setting the appropriate covariate in the heterogeneous treatment effect vector to 1 (and all others to 0)
 - Draw random confirmed-response status (given covariates, treatment assignment and current value of θ)
 - Given random confirmed response status... then draw random days to progression given simulated confirmed response status, etc...
 - “Assign” the comparator treatment and repeat from step 2.
- Draw a new θ and repeat!

Did Allowing for Heterogeneous Treatment Effects Yield Heterogeneous Results?



Discussion

- When the underlying causes of HTE are complicated (e.g., complex gene-gene or gene-environment interactions), it's possible that the best we can hope for is indirect evidence.
- Genomic evidence of HTE might not always be causal. If non-functional polymorphisms “travel along with” functional polymorphisms, they could contain significant classifying information.
 - A similar concept underlies the population genetics method of “linkage disequilibrium” analysis.

A caveat...

Moderators of Treatment Outcomes
Clinical Research and Policy Implications

Another questionable tactic is subgroup analysis. After the primary hypothesis is tested, researchers stratify the samples into men and women, different age groups, different initial severity levels, and different ethnicity groups, and then test the success of treatment separately in each such stratum. Biostatisticians have long decried this practice because of problems associated with multiple testing. The probability of a false-positive test result on 1 test might be 5%; on 2 independent tests, 10%; and on 3 independent tests, 14%. By the time 14 tests are performed (2 sex groups, 4 age groups, 3 severity levels, and 5 ethnicity levels), the probability of 1 or more false-positive test results may well exceed 50%.

A Catch-22

Can Treatment That Is Shipped on Average Be Identified as
Some Patients A Step of the Confounding Interaction Tests
of Clinical Research and Their Relevance

These data are rarely sufficient, however, for the clinician focusing on a single patient. Correct conclusion regarding treatment effectiveness for the overall study population may be incorrect for selected individuals within the population. It is natural, therefore, for the physician to engage in a process of clinical inquiry [1], assembling data that will allow assessment of the appropriate choice of treatment according to specific clinical characteristics that closely approximate those of the patient [2]. When the relevant data are derived from one or several RCTs, this, of necessity, leads to subgroup analyses. Many investigators have cautioned against performing such subgroup analyses to identify variations in treatment effects for fear of detecting differences that are attributable to chance alone [3,4]. An unresolved dilemma therefore occurs: the clinician eagerly inquiring of the data about ever more refined patient subgroups to enhance her ability to select an appropriate treatment for her patient; and the RCT trialist, strongly resisting such requests.

Concluding Thoughts

Is it time to rethink Type I error?

- “Acceptable” Type I error is tied to a specific “alpha” level, which **before** the clinical trial represents the probability that an “alternative” hypothesis is accepted when it is indeed false.
 - In the case of CEA, we might specify the null hypothesis to be that the incremental cost-effectiveness ratio of a treatment relative to comparator is \$75,000 per Quality-Adjusted Life Year (or some other critical threshold of willingness to pay per QALY), versus the alternative hypothesis that the ICER is less than \$75,000 per QALY.
- But **after** the data is collected, the “p-value” **does not** represent the probability that the null hypothesis is actually true.
- The p-value represents the proportion of a theoretically infinite number of identically conducted clinical trials in which the data obtained would generate a test statistic which, if calculated in the same way, would exceed the test-statistic actually calculated, if the null-hypothesis were true.
 - Therefore, “1 minus p-value” **does not** represent the probability that the treatment is cost-effective, given the data you’ve collected.

Hypothesis Testing vs. Decision-Making

- We’re told we can’t “accept” the null hypothesis – we just “fail to reject it.” Does that mean we should stick with the comparator treatment if our p-value is 0.051?
 - Why did we choose 0.05 anyway? Is this how decision-makers act under uncertainty?
- In economics, we typically assume that rational decision-makers facing uncertainty choose the course of action that yields the highest expected utility, given that uncertainty.
 - In CEA (or NHB), this would translate to choosing the treatment if its expected incremental net health benefit is greater than zero (see Karl Claxton, The irrelevance of inference, JHE 18(3), 1999)
 - These approaches are fundamentally Bayesian because the object of interest is the “posterior” probability of cost-effectiveness, given the data we have observed.



Photo Copyright © Rainer Mautz

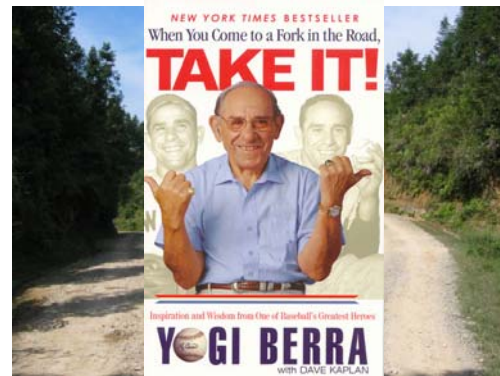


Photo Copyright © Rainer Mautz

The Bayesian Approach

- Classical “frequentist” methods give an estimate of the most likely values for θ and “confidence intervals” surrounding them.
 - But they don’t tell us what is the probability that θ equals a certain value or lies in a certain range given our data.
 - Unfortunately, this is what decision theory tells us we need!
- The Bayesian posterior is a probability distribution that summarizes our state of uncertainty about θ given our data and our model.
 - Decision analysis makes use of this concept with “probabilistic sensitivity analysis.”
- There’s no free lunch: Bayes’ Rule tells us that in order to make such a statement, we first must specify a “prior” distribution representing our beliefs about θ before observing data.
 - Such subjectivity can be troubling.
 - In this study, we used “non-informative” priors, which yield posteriors whose modes are close to the maximum likelihood estimate.
 - The ability to incorporate prior knowledge can be a strength if priors are obtained from rigorous meta-analysis methods.

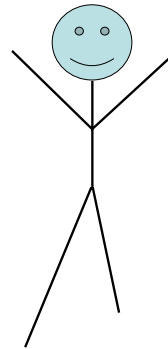
What’s Holding Us Back?

- A fear of being accused of “dredging data.”
 - Patients denied coverage because they are in a non-responsive class might feel discriminated against.
 - But: patients granted coverage because they are in a responsive class might be pleased.
 - Patients who learn that they are in a high-side-effect risk class might feel lucky.
- Developers of treatments may be concerned by potential fragmentation of their market and therefore reduced profit.
 - But: outright “failures” of treatments may be reduced – because subgroups that benefit may be identified.
- Payers might be suspicious of the motives of developers of new treatments when subgroups with high effectiveness are identified.

- Holding on to frequentist notions of probability and inference that are not consistent with decision theory.

Steps Forward

- Continue research into classification methods based on indirect evidence of "deep" classes.
- Continue progress toward adopting Bayesian Adaptive trial designs – and incorporate Bayesian analysis of HTE into the adaptive phase. If promising subgroups emerge, change recruitment mid-course.
- Pursue drug "re-discovery" by reanalyzing clinical trial data of failed or supplanted treatments with classification and Bayesian HTE methods.
- Use classification and Bayesian HTE methods in post-marketing studies and "practical" clinical trials.
- Conduct Value of Information analysis to determine whether additional trials powered to detect difference among defined classes is worth the cost from the decision-maker's point of view.
- A neutral "Institute for Comparative Effectiveness" would help as an honest broker between payors and developers of new treatments.



Thank you!