A QUANTITATIVE FRAMEWORK FOR EVIDENCE INTEGRATION

DONALD B RUBIN
HARVARD UNIVERSITY
25 OCTOBER 2017
LISBON PORTUGAL

Evidence comes from many sources that vary in critical ways

- Precise questions that are being investigated differ across studies
- Causal exposures differ
- Outcome variables change
- Settings change (e.g., in lab, in cities, urban/rural, units themselves –families, individuals)
- Characteristics of units (e.g., male/female, race, wealth, health, BMI, adults/children)

Examples From My Experience

- Anthrax vaccine and how to deliver it
  - Dose and timing of boosters
  - People, non-human primates, guinea pigs, rats
- New anti-epileptic drug – FDA to approve for kids?
  - Animal studies with adjunctive and mono therapy
  - Adults both studies, kids only adjunctive
- Etc. in many application areas
- Many social science Meta-Analyses (e.g., RR)
- Being **quantitative** helps focus discussion

Sometimes effort is termed “Meta-Analysis” (MA); Gene Glass (1976)

- Is it “Surveying and summarizing the state of current literature on a relatively homogeneous topic”?
- Or should it be “Building and extrapolating a **response surface** for the science of the topic”?
- Former more common approach – much easier too
- Latter better approach because more relevant for science and more relevant for assessing possible interventions – important for policy, former is not
- My views have remained relatively constant since proposing the latter in 1986 at a US National Academy of Science conference

"The Future of Meta Analysis" Wachter and Straf (1990), reviewed in JASA in 1991 by Gene Glass

…But the chapter that truly speaks to the future of MA, and not to its distant and recent past, is the Chapter by Donald B. Rubin entitled "A New Perspective."

[It] is easily put into words: MA should be conceptualized as building response surfaces, not as surveying the literature. The literature, published or not on any topic, is a huge unbalanced survey …describing it statistically can result in little more than a description of research customs and habits. Rubin directs our attention away from "the literature" … and toward the scientific understanding of the phenomena…

Rubin's chapter is the sort of advice that I should have liked to have written.

Response Surface = Causal Effects as a Function of Inputs

- Inputs are scientific variables (e.g., temp, age)
- Response surface must be estimated from observed studies and so is a function of both:
  - True response surface=Science, including non-humans
  - Design factors (sample size, quality of study, publication bias, etc.,) – noise to be eliminated
  - Both Science and Design require extrapolation to obtain **response surface for humans**
- Example, effect of air pollution on humans:

Mortality(high air pollution)-Mortality(low air pollution) = f(Science [e.g., urban/rural, sex, age, race] AND Design [e.g., random/obs, animals closeness to humans]

The Future of Meta-Analysis.
Kenneth W. Wachter and Miron L. Straf (eds.). New York: Russell Sage Foundation, 1990. xxvii + 210 pp. $29.95.

Wachter and Straf's book is a collection of papers, constituting 16 chapters in published form, that were originally given at a workshop on meta-analysis (MA) held by the Committee on National Statistics of the National Research Council in October 1986. In a sense, with a half-decade lag in reaching print, the future can be said to be now.

The reader will find present among the authors of these individual pieces such pioneers in data synthesis as Ingram Olkin, Larry Hedges, Norman Bradburn, Robert Rosenthal, and Fred Mosteller. Judging from the book, the workshop was better run than most and its organizers succeeded in focusing attention on a few common concerns. In large part this success can be credited to the use of two examples of MA that participants studied and prepared commentary on before the workshop; one example was an MA of about a dozen experiments on the treatment of aphasia, and the other example was several MA's by different individuals of the same set of studies on the effects of school desegregation on Black students' achievement. The authors of the current work had in hand the results of the meta-analysts' efforts. Not surprisingly, perhaps, the revisiting of the multiple MA's prompts more reflections on the relationship of empirical research to policy and on the psychology of belief than it inspires new insights into research synthesis.

Three chapters are likely to be of particular interest to *JASA* readers. Olkin traces the history of classical statistics' attention to the unique problems arising from the existence of multiple studies of a research question. This attention was limited to questions of power and statistical significance: Do six, only marginally significant, results support rejection of the null hypothesis when taken in toto? Olkin's brief recounting is marked by brilliant felicities of style: "The exclusive use of p-values has become a disease. It is fostered by journals and by granting agencies. It is not at all unusual to see one, two, or three stars in every article of most journals in the social and psychological sciences. This is a form of statistical Star Wars." (p. 5).

Larry Hedges presents a selective and intelligent review of developments in MA spurred by the renewed attention to the topic that dates from the mid-1970s. But the chapter that truly speaks to the future of MA, and not to its distant and recent past, is the chapter by Donald B. Rubin entitled "A New Perspective."

Rubin's new perspective is easily put into words: MA should be conceptualized as building and extrapolating response surfaces, not as surveying the literature. The literature, published or not, on any topic is a huge unbalanced survey; coding it, measuring it, and describing it statistically can result in little more than a description of research customs and habits. Rubin directs our attention away from "the literature" and its message and toward the scientific understanding of the phenomena and the research process by which the phenomena are explored. "... what we want to do is estimate 'true' effects of treatments ... and their interactions with scientifically important moderator variables." (p. 157) Rubin wishes to use existing studies—he raises, in passing, the seldom recognized crucial question of what a "study" is and why the social sciences have organized empirical work around this arbitrary unit—to plot the response of experimental versus control differences on output variable Y as a function of scientifically important moderator variable X and research technique variables Z. He then proposes extrapolating the surface into the region of better and better Z variables. Forget that he speaks of the point Z_0, where the "ideal study" resides. Platonism aside, his is an eminently sensible, indeed altogether wholesome, conceptualization of what research synthesis should be.

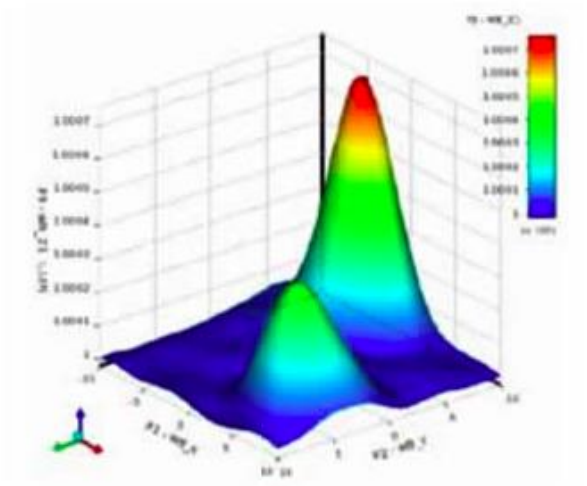
Rubin's chapter is the sort of advice I should have liked to have written. In fact, I wrote the exact opposite advice in 1983 (Glass and Kliegl 1983) when I proposed that "the literature" itself can be separated from the phenomena it is about and may itself become the object of scientific inquiry. This position is defensible; it simply is not as interesting or as important as the perspective that Rubin has now clearly articulated.

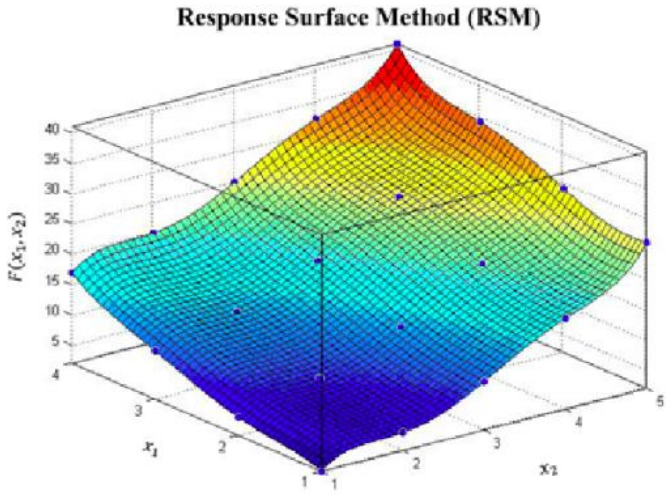
GENE V GLASS
Arizona State University

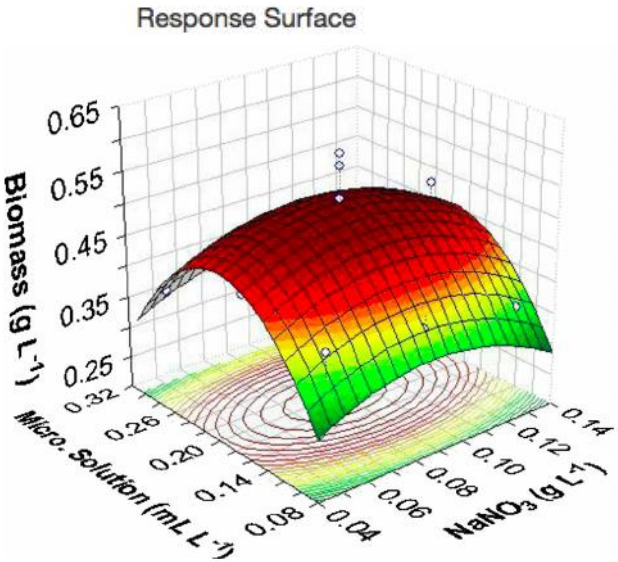
Complications with implementation of response surface conceptualization

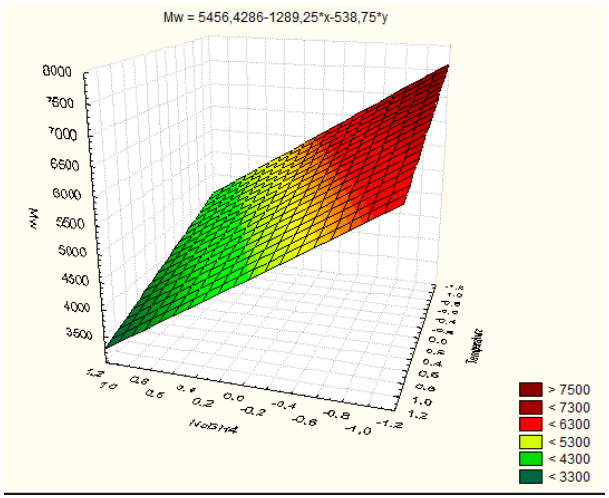
- Extrapolating response surface to “perfect study answers for humans”, from fallible actual studies
- Defining quality of various studies, especially observational (e.g., are they designed without access to outcome data? -- rarely)
- Even RCTS have complications (e.g., missing data, non-compliance, placebo versus treatment effects)
- Need to rely on models and consequential sensitivity of conclusions to assumptions presently unassailable
- Be explicit about otherwise hidden assumptions and try to justify them and investigate sensitivity to them
  - Be explicitly Bayesian!

Illustrative response surfaces









How to use a successfully estimated response surface to guide recommendations and decisions?

- Usually omitted quantitative task:
  - Formalize loss functions with public input to help achieve agreement from various sides
  - Often, there exists implicit agreement on asymmetry of loss functions but left to "politicians" to debate proposed actions without consideration of loss functions
  - Real example: CDC and in vitro fertilization

Simple example of asymmetric loss function with time as the argument

Five minutes late for airplane =>

Loss = miss flight (UGH!)

Five minutes early for airplane =>

Loss = wasted five minutes sitting at airport (so?)

More formal arguments sometime help clarify: focusing attention on where there is agreement and disagreement – USE WHEN POSSIBLE !

Addendum

- Role of sensitivity analyses
- Show, visually, how conclusions can change with varying assumptions
- E.g., at US FDA – “Tipping point” displays for missing data in experiments
- And “Enhanced tipping point displays”
- Generalize techniques to cross-species studies
- New research effort!