How I was converted from skeptic to enthusiast about meta-analysis

When I first heard of meta-analysis in the mid-1980s, I was suspicious. After slaving for 3 years on a case-control study of renal adenocarcinoma that yielded just 3 papers, the idea of someone else getting a quick publication by quantitatively pooling my findings with other studies seemed parasitic, if not plagiarist. On the other hand, I was aware that reviews of scientific literature often simply bolstered the authors' opinions by citing supportive studies and glossing over, dismissing or overlooking unsupportive studies.

My suspicion was reduced by Meir Stampfer's important paper pooling results of streptokinase trials. I decided meta-analysis might be appropriate for drug trials. However, I was troubled by the conflict between the apparent goal of meta-analysis -- to achieve statistical significance or tight confidence intervals — and Rothman's and others' criticism of p-values. Meta-analysis seemed to me to be a fancy name for combining results to get a better p-value.

In my view, p-values are pseudo-scientific, according to Karl Popper's criterion for distinguishing science from pseudoscience. Having revealed the logic of scientific discovery, Popper showed that science advances by weeding out theories, not by proving them. Consequently knowledge does not just emerge from an accumulation of facts. For example, a spontaneous new theory, without any further data collection, can suddenly produce a new interpretation of past observations and refute competing theories.

In traditional statistics, p-values are used as a metric of support for the causal hypothesis of interest. The term "testing the null hypothesis" is a surrogate for "supporting the alternative hypothesis" and has little to do with scientific testing of alternative explanations. Likewise, meta-analysis in the 1980s seemed to me to be the same sort of pseudoscience, viewing knowledge as steadily emerging from an accumulation of facts.

While I was receptive to meta-analysis of drug trials, I objected that non-experimental studies should not be meta-analyzed because "chance" is not their main source of error. Averaging relative risks across studies would merely exaggerate the pseudo-certainty by yielding a narrow confidence interval that would further understate the true error.

However, my objection was weakened by a meta-analysis of alcohol and breast cancer by a colleague, Matt Longnecker, in Tom Chalmers' meta-analysis group. They decided to weight studies not just on size but also on quality, so as to account for other biases. I was receptive to this strategy of explicitly quantifying what traditional qualitative reviewers do implicitly.

Still I was not comfortable. I had recently learned the concept of publication bias and realized I had committed such bias repeatedly by not publishing my null or ambiguous findings. I was sure that publication bias was much worse for non-experimental studies.

Sander Greenland's explanation of quantitative review of epidemiologic literature opened my mind further, because he suggested a different approach. He was not aiming for a single summary of the studies but an exploration of their differences.

He also allowed the meta-analyst to use estimated correction factors to adjust relative risks for possible biases from other causes of error.

I began wondering what a Popperian 'deductive' approach to meta-analysis would be. I had already satisfied myself (if not my readers) that such an approach leads to more rigourous epidemiologic studies. By being vigilant about hypothesizing potential alternative explanations, it forced me to design studies more carefully and test competing hypotheses more thoroughly.

The opportunity to do a Popperian meta-analysis arose in 1991. I was offered a contract to write a review of epidemiologic studies of alcohol intake and myocardial infarction (MI). I remember my visceral response to discovering the extra workload from adopting a Popperian strategy. To rule out alternative explanations, I realized I should meta-analyze not just studies of alcohol and MI, but all studies that bear indirectly on potential confounding of the relation between alcohol and MI, including studies of smoking, obesity, exercise, diet, drugs and other diseases. I made some efforts in this direction but I never managed to fully meta-analyze those ancillary studies. I discovered that a deductive meta-analysis of non-experimental studies was far from a 'quick and easy paper'.

Another Popperian epidemiologist, Charlie Poole, independently came to the conclusion that 'aggregative' meta-analysis is a misleading strategy that easily overlooks informative differences among studies. In contrast, 'explanatory' meta-analysis aims to test alternative explanations within the meta-analysis. He later told me that, while drawing this conclusion, he was unaware of the connection with Popperian thinking. I regard this as futher corroboration of my hypothesis that Popper's hypothetico-deductive method improves rigor in epidemiology.

In 1993, on a panel at the Society for Epidemiologic Research, Greenland poked holes in the practice of using subjective scores for weighting studies by quality. He challenged us to ponder whether we would accept a nutrition study in which dietary quality scores were based on lumping foods that were rated by a panel of nutritionists on a scale from "very good" to "very bad," rather than measurements of the foods' nutritional content. By analogy, he explained, we should do meta-regressions to discover study characteristics that explain differences among study findings.

I recall making a Popperian contribution to that discussion. I suggested, in defence of those who had used quality scores, that we had learned from their mistakes. The discovery of the problems of aggregate quality scores revealed a hidden fallacy in reviews that rely on the qualitative judgments of expert authors to decide what papers should be included and given importance.

During the 1990s, in various positions in the British Columbia Ministry of Health in Canada, I witnessed how top decision makers were overwhelmed with information and hurried decisions. Evidence often reached them in sound bites during seemingly chaotic meetings with packed agendas. I became more tolerant of imperfect meta-analyses because they were much better than sound bites of opinion. Ten years ago it was acceptable to flash a photocopy of a publication during a meeting and say something like "This paper shows our policy will work (or not work.)" Now accepted practice is to click the Web and retrieve the Cochrane systematic review.

In 2000, I attended the Cochrane Colloquium for the first time. I was most impressed by a presentation that might be called 'meta'-meta-analysis. It was a systematic review of systematic reviews to assess the potential biases of different

methodologies for meta-analysis. This showed that my opinions about the problems of meta-analysis are testable hypotheses and the science of meta-analysis is evolving as Popper's theory would predict.

In 2003, I joined a Cochrane group led by Andy Oxman to collaborate on systematic reviews of pharmaceutical policy impact studies. These include "laws, rules, financial and administrative orders made by governments, non-government organisations or private insurers that are intended to directly affect the use or cost of drugs." The reviews include randomised trials, non-randomised trials, interrupted time series analyses, and controlled before-after studies. That experience has further opened my mind to the potential scope of systematic reviews.

Oxman arranged for our interim findings to be assessed by an advisory committee of drug plan directors from six countries. The committee confirmed our framework for categorizing topics but requested that we put greater emphasis on modifiers of policy impacts. They wanted to know more than the average effect of a policy. They wanted to know the relative effects of different options, particularly in different policy contexts. In their own terms, they were echoing Greenland's, Poole's and my preference for explanatory rather than merely aggregative meta-analysis.

The decision makers at that workshop also were concerned that our strict inclusion criteria would exclude studies that they need to know, such as papers giving insights into different policy contexts and poor studies that are cited by critics of their policies. That led me to write an unsuccessful grant proposal to assess what methodologic principles might be needed for scientific review of unscientific studies.

My current research evaluates knowledge-translation initiatives to improve prescribing and chronic disease management. We are grappling with the conflict between validity and brevity when summarizing evidence for busy clinicians. I expect this to contribute to further evolution of my views on the evolving science of systematic reviews.

Malcolm Maclure
Professor, School of Health Information Science,
University of Victoria, British Columbia, Canada and
Adjunct Professor, Department of Epidemiology,
Harvard School of Public Health, Boston