Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-19-05

Meta-Analysis for Medical Decisions

Charles F. Manski
Board of Trustees Professor in Economics and IPR Fellow
Northwestern University

Version: February 7, 2019

DRAFT
Please do not quote or distribute without permission.

ABSTRACT

Statisticians have proposed meta-analysis to combine the findings of multiple studies of health risks or treatment response. The standard practice is to compute a weighted-average of the estimates. Yet it is not clear how to interpret a weighted average of estimates reported in disparate studies. Meta-analyses often answer this question through the lens of a random-effects model, which interprets a weighted average of estimates as an estimate of a mean parameter across a hypothetical population of studies. The relevance to medical decision making is obscure. Decision-centered research should aim to inform risk assessment and treatment for populations of patients, not populations of studies. This paper lays out principles for decision-centered meta- analysis. One first specifies a prediction of interest and next examines what each available study credibly reveals. Such analysis typically yields a set-valued prediction rather than a point prediction. Thus, one uses each study to conclude that a probability of disease, or mean treatment response, lies within a range of possibilities. Finally, one combines the available studies by computing the intersection of the set-valued predictions that they yield. To demonstrate decision- centered meta-analysis, the paper considers assessment of the effect of anti-hypertensive drugs on blood pressure.

1. Introduction

In principle, the quality of medical decision making should benefit from the proliferation of studies that analyze evidence on health risks and treatment response. In practice, difficulties arise when developers of clinical practice guidelines and individual clinicians attempt to combine findings from multiple studies. Medical decision makers must somehow interpret the mass of information provided by evidence-based research. The hard question is how to interpret this information sensibly.

Combining findings is often performed by *systematic review* of a set of studies. This is a subjective process, akin to exercise of clinical judgment. For example, evidence from dozens of randomized trials of treatments for hypertension was reviewed by the Eighth Joint National Committee (JNC 8), which developed the 2014 guidelines for management of high blood pressure in the United States (James *et al.*, 2014). The JNC 8 team wrote (p. 509):

> "An external methodology team performed the literature review, summarized data from selected papers into evidence tables, and provided a summary of the evidence. From this evidence review, the panel crafted evidence statements and voted on agreement or disagreement with each statement. For approved evidence statements, the panel then voted on the quality of the evidence."

Statisticians have proposed *meta-analysis* in an attempt to provide a more objective methodology for combining the findings of multiple studies. Meta-analysis is easy to motivate when combining findings poses a purely statistical problem. Suppose that one wants to estimate as precisely as possible some parameter characterizing a study population. The parameter may, for example, be the probability that a member of a patient population will develop a disease or mean response to a specified treatment.

Suppose that multiple studies have been performed, each analyzing an independent random sample from the population. If the raw outcome data were available, the most precise way to estimate the parameter would be to combine the samples and compute the estimate using all the data. Often, however, the raw data are unavailable, making it infeasible to combine the samples. Instead, multiple parameter estimates may be available, each computed with the data from a different random sample. Meta-analysis proposes methods

to combine the multiple estimates. The standard proposal is to compute a weighted-average of the estimates, the weights varying with sample size to minimize variance.

Meta-analysis as described above is uncontroversial, but its applicability is very limited. It is rare to have multiple random samples drawn from the same population. It is common for multiple studies to be performed with different sampling processes. The available studies may examine distinct patient populations whose members may have different probabilities of diseases or different distributions of treatment response. The protocols for administration of treatments and the measurement of outcomes may vary across studies. Meta-analyses are performed regularly in such settings, computing weighted averages of estimates for distinct study designs and study populations. Gene Glass, who introduced the term *meta-analysis*, recognized the challenge of combining findings from disparate studies. He wrote (Glass, 1977, p. 358): “The tough intellectual work in many applied fields is to make incommensurables commensurable, in short, to compare apples and oranges.”

A severe deficiency of the prevailing practice of meta-analysis is that it is not clear how to interpret a weighted average of the estimates reported in disparate studies. Meta-analyses often answer this question through the lens of a random-effects model, as proposed by DerSimonian and Laird (1986). A random-effects model assumes that the finding reported in each study estimates a distinct parameter value drawn at random from a population of potential parameter values. A weighted average of the estimates is interpreted to be an estimate of the mean parameter value across studies.

The relevance to medical decision making of the mean parameter value across studies is obscure. DerSimonian and Laird consider each of the studies considered in a meta-analysis to be drawn at random (p. 181) “from a population of possible studies.” They thus interpret the weighted average of estimates computed in the meta-analysis as an estimate of the mean outcome across the postulated population of possible studies. However, they do not explain what is meant by a population of possible studies, nor why the published studies should be considered a random sample from this population. Even if these concepts are meaningful, they do not explain how a mean health outcome across a population of possible studies

relates to what should matter to a clinician or guideline developer, namely the mean outcome across a relevant population of patients.

Researchers schooled in the paradigm of meta-analysis sometimes use *meta-regressions* to describe how study findings vary with observed attributes of the studies. Stanley and Jarrell (1989), Thompson and Higgins (2002), and Tipton, Pustejovksy, and Ahmadi (2018) provide some perspectives. Meta-regressions essentially perform meta-analysis within sub-populations of studies with specified attributes. Nonparametric meta-regressions compute weighted averages of estimates within sub-populations of studies, and parametric ones fit study findings to parametric regression models. Either way, the objective is to characterize how mean outcomes vary across sub-populations of studies. Meta-regressions do not characterize how mean outcomes vary across sub-populations of patients who vary in their attributes.

From the perspective of medical decision making, the fundamental problem of meta-analysis (and by extension meta-regression) is that the research is study-centered rather than decision-centered. Decision-centered research should aim to inform risk assessment and treatment choice for populations of patients, not populations of studies. Weighted averages of study estimates generally lack a clear interpretation from a decision perspective. Decision makers who misinterpret them as pertaining to their patient populations may be led astray. Section 2 adds to this critique of study-centered meta-analysis.

A clinician or guideline developer who wants to use existing medical research appropriately must understand how both statistical imprecision and identification problems limit the information that studies provide. Statistical imprecision is the problem of drawing inferences about a study population by observing a finite sample of its members. The severity of the problem diminishes with sample size.

Identification problems are the inferential difficulties that persist when sample size grows without bound. They encompass the spectrum of issues that decision makers have in mind when they express concern about the *internal* and *external validity* of studies. A prominent concern with internal validity is proper interpretation of observational studies of treatment response in the absence of knowledge of the process of treatment selection. A prominent concern with external validity is proper extrapolation of findings in randomized trials from study populations of trial subjects to relevant patient populations.

Manski (2018, 2019) provide extensive discussion of identification problems in evidence-based medical research.

Identification problems commonly are the dominant difficulties which arise in medical decision making. Yet the prevailing methodology of study-centered meta-analysis adequately addresses only statistical imprecision. Meta-analyses acknowledge identification problems only when they divide existing studies into those that will and will not be included in the analysis. Studies thought to have severely limited internal or external validity may be excluded from consideration.

If meta-analysis were to be decision-centered rather than study-centered, it would address identification problems prominently and credibly. In general, one would not compute a weighted average of study estimates. The methodology developed in this paper instead computes the intersection of set-valued predictions derived from the available studies.

Section 3 lays out principles for decision-centered meta-analysis. To begin, one specifies a prediction of interest---designating the relevant patient population and the health outcome to be predicted. One next determines what each available study credibly reveals about this outcome. Analysis of identification with credible assumptions typically yields a set-valued rather than point prediction. That is, one uses an existing study to credibly conclude that a probability of disease or mean treatment response lies within a range of possibilities.

The final step of decision-centered meta-analysis as developed here is to combine the conclusions derived from the available studies by computing the intersection of the set-valued predictions that they yield. This intersection expresses what one can learn by combining multiple studies. The prediction of interest must simultaneously lie in the set-valued prediction obtained with each study separately.

For example, consider combining findings from an observational study and a randomized trial. Jointly considering internal and external validity, an observational study and a trial may each suffice to bound but not point-identify a prediction of interest. The truth must lie in the intersection of the bounds obtained with each type of data.

The use of set intersection rather than a weighted average to aggregate information across studies may appear radical from the perspective of study-centered meta-analysis. However, the idea has precedent in the econometrics literature on partial identification, from Manski (1990) onward. See Section 3.4 for discussion.

Section 4 provides a simple demonstration of decision-centered meta-analysis. When developing the 2014 guidelines for management of high blood pressure in the United States (James *et al.*, 2014), the JNC 8 group completely dismissed findings from observational studies. It considered only findings from randomized trials. When doing so, the committee gave much attention to trials whose study populations differ substantially from the relevant American population. I sketch how a decision-centered meta-analysis might combine findings from these trials and from a publicly available American observational study, the National Health and Nutrition Examination Survey (NHANES).

2. Critique of Study-Centered Meta-Analysis

The technical features of study-centered meta-analysis and meta-regression have been described in many sources; see, for example, Glass (1977), Hedges and Olkin (1985), Thompson and Higgins (2002), and Tipton *et al.* (2018). There is no need to repeat such description here. There is, however, reason to elaborate on my conceptual criticism of the central meta-analytic idea of using a weighted average of estimates to aggregate the information in a population of studies. The curious nature of this idea has been remarked on only occasionally in the literature on meta-analysis.

Section 2.1 adds to my earlier discussion of random-effects models. Section 2.2 criticizes occasional suggestions that averaging estimates is justified by the so-called “wisdom of crowds.” Section 2.3 calls attention to the relatively rare cases in which medical researchers present a range of estimates rather than an average. Although motivated only heuristically, this is a positive departure from the norm in meta-analysis.

2.1. Random-Effects Models

Medical researchers have used random-effects models to perform numerous meta-analyses of studies evaluating treatments for many diseases. To illustrate, consider the highly cited article of Buchwald *et al.* (2004), who combined the findings of 134 studies of the outcomes of bariatric surgery. The 134 studies included 5 randomized trials enrolling 621 patients and 28 nonrandomized but somehow otherwise controlled trials enrolling 4,613 patients. The 101 other studies, described as "uncontrolled case series," involved 16,860 patients. The studies were performed around the world, 58 with European patients, 56 with North American patients, and 20 with patients from elsewhere. The studies followed patients for different periods of time. They measured weight loss, a primary health outcome of interest, in multiple ways.

To summarize the findings of the meta-analysis, the authors wrote (p. 1724): "The mean . . . . percentage of excess weight loss was 61.2% . . . . for all patients." The mean value of 61.2% considers the 134 studies to be a random sample drawn from a population of potential studies. It is not clear what implications should be drawn by a clinician who treats a population of patients, not a population of studies.

Medical researchers who have performed meta-analyses using random-effects models have struggled to explain how clinicians should use the findings. Consider, for example, the meta-analysis performed by Chen and Parmigiani (2007) of ten disparate studies predicting risk of breast and ovarian cancer among women who carry BRCA mutations. The authors describe a weighted average of the risks reported by all studies as a (p. 1331) "consensus estimate." However, there is no consensus across the studies, which reported heterogeneous estimates pertaining to heterogeneous study populations. Chen and Parmigiani concluded their article with this guidance to clinicians (p. 1333):

> "In current clinical practice, two scenarios are possible. In the first, the clinician is able to identify a single study that matches the relevant patient population for his or her practice. In the second, which is perhaps more common, there is no clear criterion for deciding which study is most appropriate for a particular patient. In this case, given current knowledge, a meta-analysis that

> acknowledges heterogeneity is the most evidence-based and, arguably, ethically sound approach to risk counseling."

Although meta-analysis is "evidence-based," this descriptor should not reassure clinicians. What matters is whether a methodology uses evidence reasonably.

One may ask how DerSimonian and Laird, whose 1986 article proposed use of the random-effects model in meta-analysis, motivate medical use of the model. In a recent retrospective article, they acknowledge but belittle criticism of the idea of a random sample of studies, writing (DerSimonian and Laird, 2015, p. 142):

> "An early criticism of the method is that the studies are not a random sample from a recognizable population. As discussed in Laird and Mosteller [28], absence of a sampling frame to draw a random sample is a ubiquitous problem in scientific research in most fields, and so should not be considered as a special problem unique to meta-analysis. For example, most investigators treat patients enrolled in a study as a random sample from some population of patients, or clinics in a study as a random sample from a population of clinics and they want to make inferences about the population and not the particular set of patients or clinics. This criticism does not detract from the utility of the random-effects method. If the results of different research programs all yield similar results, there would not be great interest in a meta-analysis. We view the primary purpose of meta-analysis as providing an overall summary of what has been learned, as well as a quantitative measure of how results differ, above and beyond sampling error."

This statement expresses the essence of the motivation for study-centered meta-analysis. It does not persuasively justify the use of meta-analysis in medical decision making.

2.2. Averaging and the "Wisdom of Crowds"

A proponent of meta-analysis might suggest that averaging the findings of multiple studies tends to work well empirically, even though the practice may be deficient logically. To support this suggestion, the proponent might refer to the "wisdom of crowds."

The wisdom of crowds is the name given by Surowiecki (2004) to an often-reported empirical regularity. Empirical researchers who study prediction in various fields of science have long found that the

average of a set of predictions tends to be more accurate than the individual predictions used to form the average. Formally, they report that the prediction error of the average prediction is smaller than the average error across the individual predictions. A review article by Clemen (1989) put it this way (p. 559):

> "The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy. This has been the result whether the forecasts are judgmental or statistical, econometric or extrapolation. Furthermore, in many cases one can make dramatic performance improvements by simply averaging the forecasts."

Citing this literature, one may be tempted to hope that, however deficient the logic of meta-analysis may be, the methodology may work in practice. I strongly caution against this.

McNees (1992) and Manski (2011, 2016) show that the wisdom of crowds is not an empirical regularity. It is rather an algebraic result that holds whenever the loss function used to measure prediction error is such that one suffers an increasing marginal loss as the magnitude of the error grows. In mathematical terms, such a loss function is convex. Prominent examples of convex loss functions, used routinely in studies of predictive accuracy, are square and absolute loss.

With a convex loss function, the result known as the wisdom of crowds holds as a result of Jensen's Inequality. It holds regardless of whether one combines predictions by their simple average or by a weighted average. Being an algebraic property rather than an empirical regularity, the result holds regardless of the quality of the individual predictions that are combined. Thus, although averaging predictions tends to perform well relative to the individual predictions used to form the average, it need not perform well in an absolute sense. It may provide a good prediction or a terrible one in absolute terms.

2.3. Suggestions to Report Ranges rather than Averages of Estimates

Although averaging predictions across studies has been the norm in medical meta-analysis, investigators occasionally present a range of predictions and describe the range as measuring uncertainty. I use breast-cancer risk assessment to illustrate.

The Gail Model, developed in Gail *et al.* (1989), is perhaps the most prominent model that predicts future development of breast cancer in women with specified attributes. However, it is not the only such model. A review article of Amir *et al.* (2010) considers five other models as well: the Claus Model, the BRCAPRO Model, the Jonker Model, the IBIS Model, and the BOADICEA Model. These five studies differ from one another in various respects, including the patient attributes used to condition predictions, the mathematical forms of the models, and the data used to estimate model parameters. As a result, they yield different probabilistic predictions when applied to women with specified attributes.

Domchek *et al.* (2003) compare the Gail and Claus Models and find that (p. 597): "Concordance of the two models is only fair." Domchek *et al.* reject the notion that a clinician should use any weighted average of the predictions of the two models. Instead, they suggest that these models (p. 600) "may provide helpful ranges" of probabilistic predictions.

Going further, Mandelblatt *et al.* (2009) report on an ambitious project that uses multiple models to generate a range of hypothetical predictions of breast cancer development and mortality under alternative strategies for mammography screening. The authors write (p. 740):

> "Each model has a different structure and assumptions and some varying input variables, so no single method can be used to validate results against an external gold standard. . . . Overall, using 6 models to project a range of plausible screening outcomes provides implicit cross-validation, with the range of results from the models as a measure of uncertainty."

Domchek *et al.* (2003) and Mandelblatt *et al.* (2009) appropriately recognize that, when multiple plausible models generate a range of predictions, the correct prediction is uncertain. They suggest using the range of results to quantify the uncertainty. The reasoning that these authors give is heuristic rather than grounded in formal analysis. Nevertheless, presentation of a range of predictions is at least a step in a positive direction away from averaging.

3. Decision-Centered Meta-Analysis for Personalized Patient Care

I now develop principles for decision-centered meta-analysis. I first describe a clinician's prediction objectives. I then consider use of study findings to partially achieve the objectives.

3.1. Clinical Prediction Objectives

Consider a clinician who wants to personalize patient care. One might ideally think of personalized care as literally specific to the individual patient, but knowledge to support complete personalization is generally not available. Personalized care in practice means care that varies with clinically observed patient attributes.

Suppose that a clinician wants to probabilistically assess health risks or predict treatment response conditional on observed patient attributes. Existing evidence-based studies may provide some relevant information, but they may not yield fully accurate probabilistic predictions. The question is how a clinician may reasonably use the available study findings.

To formalize, let there exist a population of patients and let the particular patient under consideration be denoted as patient "0." Let x_0 denote the observed patient attributes. Let y denote a patient health outcome of clinical concern.

Probabilistic risk assessment means that the clinician wants to know the conditional probability distribution $P(y|x_0)$. When predicting treatment response, let T denote a set of alternative possible treatments. For each $t \in T$, let y(t) denote the health outcome that the patient would experience with treatment t. Probabilistic prediction of treatment response means that the clinician wants to know the conditional distributions $P[y(t)|x_0]$, $t \in T$. It is essential to recognize that the probability distributions under discussion here are distributions of outcomes among patients with attributes x_0. They are not distributions across some conjectured population of studies.

Clinicians often concentrate attention on mean health outcomes rather than complete distributions. Then the quantity of interest may be $E(y|x_0)$ or $E[y(t)|x_0]$, $t \in T$. I focus on these quantities henceforth.

3.2. The Information Provided by Studies

Existing studies may not fully reveal the mean outcomes that the clinician wants to know. Suppose that the findings of a set K of studies have been reported. Each was conducted in some study population, measuring study-specific outcomes and patient attributes. In studies of treatment response, the treatments administered may have been study-specific as well. Thus, study k may report an estimate of a conditional mean risk-assessment $E(y_k|x_k)$ or of mean treatment-responses $E_k[y_k(t)|x_k]$, $t \in T_k$.

The findings reported in study k may differ from the one of clinical interest for several reasons. Outcome y_k differs from y if the study measured a surrogate outcome, whereas y is an outcome of real health interest. Patient attributes x_k differ from x_0 if the study had subject inclusion criteria that excluded patients with attributes x_0 from participation. If study k was a trial, the treatments in T_k differ from the clinically relevant treatments T if subjects in the trial received blinded treatments or if they received heightened attention relative to what patients receive in clinical practice.

While the reported study findings do not pin down the mean outcomes of clinical interest, they may be informative. To focus on identification, suppose that the study sample was large enough to make statistical imprecision a negligible concern. Thus, consider the estimates of $E(y_k|x_k)$ or $E_k[y_k(t)|x_k]$, $t \in T_k$ to be accurate.

Knowledge of these quantities, together with other information provided by the study investigators and with credible assumptions, may enable the clinician to conclude that the mean outcomes of interest lie within certain sets of possibilities. Formally, one may be able to conclude that $E(y|x_0)$ lies within some set E_{k0} on the real line or that $E[y(t)|x_0]$, $t \in T$ lies within some set H_{k0} of dimension $|T|$.

In many applied settings, the set E_{k0} takes the form of an interval. Thus, one may be able to conclude that $E(y|x_0) \in [e_{k0L}, e_{k0U}]$, where e_{K0L} and e_{K0U} are the lower and upper bounds of the interval. Similarly, one

may be able to conclude that, for each $t \in T$, $E[y(t)|x_0]$ lies in an interval $[e(t)_{k0L}, e(t)_{k0U}]$. Section 3.4 will give specific forms that these intervals may take in practice.

In each case, the width of the interval measures the informativeness of study k to the clinician. The study is fully informative when the lower and upper bounds of the interval coincide. It is uninformative when the lower and upper bounds are the smallest and largest logically possible value of outcome y_0.

3.3. Aggregating Information across Studies

Section 3.2 considered one study in isolation. Meta-analysis aggregates information across multiple studies. Consider risk assessment. When the existing studies partially identify $E(y|x_0)$, the appropriate way to aggregate information across studies is to form the set intersection $\cap_{k \in K} E_{k0}$. A potential value of $E(y|x_0)$ is consistent with all the information in the studies if and only if the value lies within the set intersection. Similarly, a potential value of the mean treatment-response vector $E[y(t)|x_0]$, $t \in T$ is consistent with the information in all studies if and only if it lies within the set intersection $\cap_{k \in K} H_{k0}$.

Intersection of sets takes a simple form when the sets are intervals. Then the intersection is itself an interval, whose lower bound is the greatest lower bound of the study-specific intervals and whose upper bound is the least upper bound of these intervals. For risk assessment, the set intersection is the interval $[\max_{k \in K} e_{k0L}, \min_{k \in K} e_{k0U}]$. For mean treatment response, it is $[\max_{k \in K} e(t)_{k0L}, \min_{k \in K} e(t)_{k0U}]$.

The operation of set intersection differs markedly from the traditional meta-analytic practice of computing a weighted average of estimates. Depending on the applied context, a meta-analytic weighted average risk assessment may or may not lie within the set intersection $\cap_{k \in K} E_{k0}$. The same remark holds for mean treatment response. When the evidence and assumptions used to compute the set intersection are accurate, failure of a weighted-average estimate to lie within this set implies that the weighted average cannot correctly predict the quantity of clinical interest. Thus, the hypothesis that a traditional meta-analytic weighted average provides a correct prediction may be refutable.

The above discussion supposes that the evidence and assumptions used to compute the set intersection are accurate. It may be that some study investigators misinterpreted their evidence or that the clinician performing decision-centered meta-analysis imposed inaccurate assumptions when interpreting the study findings. In some applied settings, the operation of set-intersection may then yield an empty set. If this occurs, one should conclude that some study findings or clinician assumptions are incorrect. Thus, the hypothesis that a decision-centered meta-analysis provides a correct set prediction may be refutable.

3.4. Analysis of Set Intersection in Research on Partial Identification

The use of set intersection to aggregate information across studies has precedent in econometric research on partial identification. I summarize here.

Manski (1990) derived nonparametric bounds on mean treatment response obtained from observational studies, where one may have no knowledge of the process of treatment selection. Let outcome y(t) have known bounded range, say $[y_L, y_U]$. Consider a set K of observational studies of populations that are credibly thought to have the same distribution of treatment response as the patient population of interest. Let each study population contain some persons who have attributes x_0. Let each study measure outcome y(t) for those members of the study population who receive treatment t. Let z_k denote the treatment received by a member of study population k.

Let $P(z_k = t|x_0)$ be the fraction of persons in study population k who receive treatment t, among those with attributes x_0. Let $E[y(t)|x_0, z = t]$ be mean treatment response within the group who have attributes x_0 and who receive treatment t. The quantities $P(z_k = t|x_0)$ and $E[y(t)|x_0, z_k = t]$ are in principle observable to an investigator performing observational study k. On the other hand, the quantity $E[y(t)|x_0, z_k \neq t]$ is counterfactual, hence unobservable.

The Law of Iterated Expectations relates the above quantities to $E[y(t)|x_0]$ as follows:

$$(1) \quad E[y(t)|x_0] = E[y(t)|x_0, z_k = t]\cdot P(z_k = t|x_0) + E[y(t)|x_0, z_k \neq t]\cdot P(z_k \neq t|x_0).$$

With $E[y(t)|x_0, z_k \neq t]$ unobservable, we can conclude that $E[y(t)|x_0]$ lies in the interval

$$(2) \quad E[y(t)|x_0, z_k = t]\cdot P(z_k = t|x_0) + y_L\cdot P(z_k \neq t|x_0) \leq E[y(t)|x_0] \leq E[y(t)|x_0, z_k = t]\cdot P(z_k = t|x_0) + y_U\cdot P(z_k \neq t|x_0).$$

The lower and upper bounds are obtained by inserting the polar possibilities for $E[y(t)|x_0, z \neq t]$, which are that it equals y_L or y_U respectively.

Now consider the set K of studies. $E[y(t)|x_0]$ lies in interval (2) for every $k \in K$. Hence, it lies within the intersection of the $|K|$ study-specific intervals. The result is the intersection interval

$$(3) \quad \max_{k \in K} E[y(t)|x_0, z_k = t]\cdot P(z_k = t|x_0) + y_L\cdot P(z_k \neq t|x_0) \leq E[y(t)|x_0] \leq \min_{k \in K} E[y(t)|x_0, z_k = t]\cdot P(z_k = t|x_0) + y_U\cdot P(z_k \neq t|x_0).$$

Inequality (3) is a special case of the class of instrumental-variable bounds introduced in Manski (1990) and subsequently developed further in Manski (2003, 2007) and Manski and Pepper (2000, 2009). The terminology in these sources is distant from that used in the literature on study-centered meta-analysis. Hence, it is understandable if a researcher who performs meta-analysis but is not familiar with research on partial identification might not immediately recognize the applicability of the findings.

Two analyses of partial identification that overtly relate to meta-analysis are Manski (2003, 2004a). Manski (2003, Section 1.4) characterizes abstract set intersection when multiple studies use different sampling processes to examine different sub-populations of a population of interest. The findings reported there apply directly to risk assessment. Manski (2004a) combines findings on treatment response obtained from observational studies of multiple successive cohorts of persons. The findings reported there extend to dynamic settings (3) and more abstract findings on set intersection for distributions of treatment response.

Observation of an increasing number of cohorts enables one to intersect an increasing number of sets over time, thereby increasing knowledge of the distribution of treatment response as time passes.

3.5. Aggregating Findings of Statistically Imprecise Studies

To focus on the identification problems that often dominate use of evidence-based research in medical decision making, I have supposed that the available studies analyze data samples that are large enough to make statistical imprecision negligible. In practice, some samples may be small enough that imprecision is a serious concern.

A standard approach to characterization of imprecision has been to compute confidence sets for quantities that are estimated imprecisely. The literature on inference on partially identified quantities has developed several approaches to computation of confidence sets for set intersections. Some are based on simple heuristics (Kreider and Pepper, 2007) and others are grounded in formal asymptotic statistical theory (Chernozhukov, Lee, and Rosen, 2013). These approaches may be applied to decision-centered meta-analysis.

Ideally, decision-centered meta-analysis should bring to bear statistical decision theory, which aims to provide reasonable prescriptions for decision making with sample data (Wald, 1950). This is a topic for future research.

Early applications of statistical decision theory investigated best point prediction under square loss rather than treatment choice. See, for example, Hodges and Lehman (1950). The early literature on best point prediction appears not to have addressed combination of findings across studies. However, Dominitz and Manski (2017) recently do so. They consider minimization of maximum mean square error subject to a budget constraint, when multiple available studies use sampling processes that differ in cost and the data quality they yield.

Treatment choice has long been a topic of study in the Bayesian branch of statistical decision theory. See, for example, Canner (1970), Spiegelhalter, Freedman, and Parmar (1994), Cheng, Su, and

Berry (2003), and Spiegelhalter (2004). To combine findings across studies, some Bayesian statisticians have proposed *Bayesian averaging* (e.g., Hoeting *et al.*, 1999). This approach computes a weighted average of estimates, the weights expressing both statistical imprecision and subjective considerations.

Bayesian averaging is essentially a Bayesian version of study-centered meta-analysis. The approach assumes that the quantity estimated in at least one of the available studies appropriately estimates the quantity of clinical interest. It places a prior subjective probability on each study being appropriate in this sense. Bayesian averaging does not recognize that the quantity of clinical interest may differ from those estimated in all the available studies.

Appropriate recognition of this basic problem is possible in principle with application of the branch of statistical decision theory that aims to achieve uniformly satisfactory treatment choice, as measured by maximum regret across states of nature. However, the small extant literature of this type has thus far mainly sought to cope with statistical imprecision alone, in the absence of identification problems. See Manski (2004b), Stoye (2009), Manski and Tetenov (2016), and Kitagawa and Tetenov (2018). Only Stoye (2012) has considered decision making when one faces both statistical imprecision and an identification problem. Stoye examines a setting in which only one study is available. Thus, application of statistical decision theory to decision-centered meta-analysis is presently an open subject.

4. Illustration: Combining Findings on Drug Treatment of Hypertension

This section demonstrates decision-centered meta-analysis using set intersection to combine findings from randomized trials and observational studies. I keep the illustration simple to make the basic ideas plain. In practice, I anticipate that decision-centered meta-analysis would combine findings from a larger set of studies. Such analysis should bring to bear the expertise of medical researchers and clinicians with experience in risk assessment and treatment.

The matter to be considered is determination of the effect of drug treatment on the blood pressure of persons who have been diagnosed with hypertension. Systematic reviews and study-centered meta-analyses have examined this and related questions. I take as my starting point the evidence review performed to support the 2014 guidelines for management of high blood pressure in the United States, mentioned in the Introduction. I then use data from the NHANES Survey.

4.1. Evidence Review by the JNC 8 Group

The JNC 8 guideline development group completely dismissed findings from observational studies. It considered only findings from randomized trials. The group wrote (James *et al.*, 2014, p. 508): "The panel limited its evidence review to RCTs because they are less subject to bias than other study designs."

The JNC 8 performed a systematic review of the available trial evidence rather than a meta-analysis. This review highlighted three trials, described in Beckett *et al.* (2008), SHEP Cooperative Research Group (1991), and Staessen *et al.* (1997). It is instructive to juxtapose the study populations of these trials.

The participants differed in age, in countries of residence, and in the rules governing exclusions for comorbidities. Moreover, they differed in their blood pressure levels measured at two stages before being randomized into treatment. The SHEP and Staessen trials, but not the Beckett trial, restricted eligibility to persons diagnosed as having *isolated systolic hypertension*, a condition in which systolic blood pressure (SBP) is higher than desirable but diastolic blood pressure is in the normal range. I provide some details of the study populations below, drawing on descriptions in the published articles.

Beckett trial: Participants were 80 years of age or older, residing in Europe, China, Australasia, and Tunisia. Persons with various comorbidities were excluded. To be eligible, a person initially had to have SBP of 160 mm Hg or more. This group included persons who were and were not receiving antihypertensive treatment at the time. Persons with blood pressure in the eligible initial range consented to stop treatment for several

months, after which they were required to have SBP in the range 160-199 and diastolic blood pressure below 110. Persons in this subgroup were randomized to a drug treatment or to placebo.

SHEP trial: Participants were 60 years of age or older, residing in the United States. Persons with various comorbidities were excluded. To be eligible, a person initially had to have SBP in a certain range, the specifics depending on whether they were or were not receiving antihypertensive treatment at the time. Persons with blood pressure in the eligible initial range consented to stop treatment for several months, after which they were required to have SBP in the range 160-219 and diastolic blood pressure below 90. Persons in this subgroup were randomized to a drug treatment or to placebo.

Staessen trial: Participants were 60 years of age or older, residing in Europe. Persons with various comorbidities were excluded. The published article does not state whether a person initially had to have SBP in some range. Potential subjects consented to stop treatment for several months, after which they were required to have SBP in the range 160-219 and diastolic blood pressure below 95. Persons in this subgroup were randomized to a drug treatment or to placebo.

The designated primary outcome studied in each trial was the occurrence of stroke in subjects. The investigators also reported blood pressure outcomes. For example, they gave findings on mean SBP two years after initiation of the trial. At that time, means for treated and placebo subjects were 145 and 160 in the Beckett trial (Beckett *et al.*, 2008, Figure 2), 142 and 154 in the SHEP trial (SHEP Cooperative Research Group, 1991, Table 3), and 151 and 161 in the Staessen trial (Staessen *et al.* 1997, Figure 3).

4.2. NHANES Evidence on Drug Treatment and Blood Pressure

The NHANES Survey (National Center for Health Statistics, 2019) assesses the health status of the American population through a continuous cross-sectional survey whose findings are released to the public

biannually. I consider the biannual reports for the periods 2007-8, 2009-10, 2011-2012, 2013-2014, and 2015-16. Each report provides data on about 4000 respondents. The questions posed regarding hypertension and the procedure used to measure blood pressure have remained the same through these years. The sample design has remained close to the same over time. For simplicity, I examine the raw findings, ignoring the fact that NHANES oversamples some demographic subgroups.

I focus on the subsample of persons of age at least 60 (henceforth, older persons) who have at some past time been diagnosed with hypertension. Let the outcome of interest be SBP. NHANES expresses diagnosis by response to survey question BPQ020, which asks "Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?" It expresses drug treatment for hypertension by response to question BPQ050A, which asks "Are you now taking prescribed medicine for high blood pressure?" In a medical examination, NHANES staff measure each respondent's SBP at least three times, with the results reported in variables (BPXSY1, BPXSY2, BPXSY3). I use the average of these measurements.

Table 1: Average Systolic Blood Pressure of Older NHANES Respondents Diagnosed with Hypertension

period	treated	Average SBP	number responses	fraction responses
2007-16	no	142.2	305	0.06
	yes	135.3	4532	0.94
2007-8	no	145.6	58	0.06
	yes	135.9	922	0.94
2009-10	no	141.4	58	0.05
	yes	133.9	1004	0.95
2011-12	no	144.0	54	0.06
	yes	135.9	845	0.94
2013-14	no	141.6	64	0.07
	yes	134.4	858	0.93
2015-16	no	139.1	71	0.07
	yes	136.6	903	0.93

Table 1 gives the findings on average SBP for the 4837 subsample respondents across the 2007-2016 period. The table also gives the findings in each of the five biannual NHANES reporting periods. Across the ten-year period, about 94 percent of the subsample were treated and 6 percent untreated. Average SBP was 135.3 for treated persons and 142.2 for untreated ones.

If one were to assume that treatment selection is random and ignore statistical imprecision, one would conclude that mean SBP would be 135.3 if all older persons diagnosed with hypertension were treated with prescribed medicines and would be 142.2 if none were treated. The NHANES being an observational study, one may not find it credible to assume that treatment is random. The bound stated in (2) makes no assumption about the process of treatment selection. Computation of the bound only requires credible values for y_L or y_U, lower and upper bounds on the value of counterfactual mean SBP. For this purpose, I shall use the observed 2.5 and 97.5 percentile values of SBP among the 4837 NHANES subsample members, namely 100.7 and 182.0. Using these values as y_L and y_U, computation of bound (2) yields these bounds on mean SBP with and without treatment: [133.4, 138.2] and [103.2, 179.6].

These bounds show that the NHANES observational study is very informative about the mean SBP that would occur if all persons with hypertension were to be treated, placing it in the narrow bound [133.4, 138.2] without making assumptions about treatment selection. On the other hand, the survey reveals little about the mean SBP that would occur if no one were to receive treatment, placing it in the wide bound [103.2, 179.6]. This asymmetry arises from the asymmetry between the fraction of the subsample who are treated (0.94) and who are not (0.06). Among NHANES respondents, the fraction who have counterfactual outcomes with treatment is correspondingly small (0.06) and without treatment is large (0.94).

The bounds given here can in principle be narrowed further. For example, rather than pool all respondents across the ten-year period, one could compute separate bounds for each of the five biannual reporting periods and then compute intersection interval (3). The intersection bounds do not substantially narrow the bounds with pooled data because empirical findings are similar across reporting periods. Sampling imprecision becomes a non-trivial matter when computing bounds on mean SBP without treatment, because the biannual sample sizes without treatment are relatively small (54 to 71).

4.3. Combining the Trial and NHANES Findings

The Beckett, SHEP, and Staessen trials may have strong internal validity, but they have limited external validity when considering treatment of the American population. These trials mainly obtained data on study populations composed of non-Americans who were diagnosed with isolated systolic hypertension, who lacked various comorbidities, and who volunteered to participate. As usual in drug trials, treatment administration differed from what would occur in practice because treatment was double-blinded. Double-blinding prevents clinicians from following the common practice of sequentially prescribing alternative drugs for hypertension, trying each for a period with the objective of finding one that performs satisfactorily.

The NHANES data suffer from none of these problems of external validity. NHANES examines a broadly representative sample of the American population and it provides findings on the outcome of treatment administration as it occurs in clinical practice. On the other hand, NHANES has limited internal validity in the absence of knowledge of the process of treatment selection in practice.

The fact that the preponderance of persons diagnosed with hypertension receive treatment makes the NHANES problem of internal validity highly asymmetric. The NHANES data are almost uninformative about mean SBP in the total absence of treatment. However, they yield the narrow bound [133.4, 138.2] on mean SBP with treatment of all older Americans diagnosed with hypertension. The upper bound is noticeably smaller than all three trial estimates of mean SBP after two years, which are (145, 142, 151).

Set intersection provides a sensible way to combine the trial and NHANES findings. The JNC 8 or another guideline development group could quantify the seriousness of the trial problems of external validity by bounding the maximum difference they think credible between the mean trial outcomes and mean SBP in the American sub-population of interest. They could then intersect these bounds with those obtained from the NHANES data.

I caution that it may be challenging for a guideline development group to credibly quantify the external validity of the Beckett, SHEP, and Staessen trials. The study populations and treatment regimes in

these trials differ from those germane to developers of American clinical practice guidelines in many ways. Yet I believe it essential to meet this challenge. The prevailing practice in medical research has been to just mention problems of external validity of trials in verbal caveats. This cannot suffice.

To illustrate set intersection, suppose one finds it credible to assume that the absolute differences between SBP outcomes in the thee trials and the mean outcomes for the relevant American subpopulation are no greater than 15 mg Hg. Combining the bounds obtained from the trials and the NHANES data yields intersection bounds on mean SBP of older Americans with and without treatment, shown in Table 2.

Table 2: Bounds on Mean Systolic Blood Pressure of Older Americans Diagnosed with Hypertension

Treated	Beckett	SHEP	Staessen	NHANES	Intersection
no	[145, 175]	[139, 169]	[146, 176]	[103, 180]	[146, 169]
yes	[130, 160]	[127, 157]	[136, 166]	[133, 138]	[136, 138]

Observe the asymmetry between the intersection bounds for treated and non-treated persons, which are [136, 138] and [146, 169]. The NHANES and Staessen findings yield the upper and lower intersection bounds for treated persons. The Beckett and SHEP trials yield bounds that are supersets of the intersection and so do not add information. The SHEP and Staessen findings yield the upper and lower intersection bounds for untreated persons. The Beckett and NHANES findings yield bounds that are supersets of the intersection and so do not add information.

5. Conclusion

The standard practice in meta-analysis has been to compute a weighted-average of estimates reported in disparate studies. Random-effects models have interpreted this weighted average as an estimate of a mean parameter across a hypothetical population of studies. The relevance to medical decision making

is obscure. Decision-centered research should aim to inform risk assessment and treatment for populations of patients, not populations of studies.

This paper has laid out principles for decision-centered meta-analysis and has illustrated with assessment of the effect on blood pressure of anti-hypertensive drugs. To cope with the identification problems that regularly afflict medical research, I recommend that computation of the intersection of set-valued predictions should replace computation of weighted-averages of point estimates.

References

Amir, E., O. Freedman, B. Seruga, and D. Evans (2010), "Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models," *Journal of the National Cancer Institute*, 102, 680-691.

Buchwald, H., Y. Avidor, E. Braunwald, M. Jensen, W. Pories, K. Fahrbach, and K. Schoelles (2004), "Bariatric Surgery: A Systematic Review and Meta-analysis," *Journal of the American Medical Association*, 292, 1724-1737.

Canner, P. (1970), "Selecting One of Two Treatments When the Responses Are Dichotomous," *Journal of the American Statistical Association*, 65, 293-306.

Chen, S. and G. Parmigiani (2007), "Meta-Analysis of *BRCA1* and *BRCA2* Penetrance," *Journal of Clinical Oncology*, 25, 1329-1333.

Cheng, Y., F. Su, and D. Berry (2003), "Choosing Sample Size for a Clinical Trial Using Decision Analysis," *Biometrika*, 90, 923-936.

Chernozhukov, V., S. Lee, and A. Rosen (2013), "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667-737.

Clemen, R. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559-583.

DerSimonian, R. and N. Laird (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7, 177-188.

DerSimonian, R. and N. Laird (2015), "Meta-Analysis in Clinical Trials Revisited," *Contemporary Clinical Trials*, 45, 139-145.

Domchek, S., A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, and B. Weber (2003), "Application of Breast Cancer Risk Prediction Models in Clinical Practice," *Journal of Clinical Oncology*, 21, 593-601.

Dominitz, J. and C. Manski (2017), "More Data or Better Data? A Statistical Decision Problem," *Review of Economic Studies*, 84, 1583-1605.

Gail, M, L. Brinton, D. Byar, D. Corle, S. Green, C. Shairer, and J. Mulvihill (1989), "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually," *Journal of the National Cancer Institute*, 81,1879-86.

Glass G. (1977) "Integrating Findings: The Meta-analysis of Research," *Review of Research in Education*, 5, 351-379.

Hedges, L. and I. Olkin (1985), *Statistical Methods for Meta-Analysis*, New York: Academic Press.

Hodges, E. and E. Lehmann (1950), "Some Problems in Minimax Point Estimation," *Annals of Mathematical Statistics*, 21, 182-197.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382-417.

James, P, S. Oparil, B. Carter, W. Cushman, C. Dennison-Himmelfarb, J. Handler, D. Lackland, M. LeFevre, T. MacKenzie, O. Ogedegbe, S. Smith Jr, L. Svetkey, S. Taler, R. Townsend, J.Wright Jr, A. Narva, and E. Ortiz (2014), "Evidence-Based Guideline for the Management of High Blood Pressure in Adults Report From the Panel Members Appointed to the Eighth Joint National Committee (JNC 8)," *Journal of the American Medical Association*, 311, 507-520.

Kitagawa, T. and A. Tetenov (2018), "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591-616.

Kreider, B. and J. Pepper (2007), "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432–441.

Mandelblatt, J., K. Cronin, S. Bailey, D. Berry, H. de Koning, G. Draisma, H. Huang, S. Lee, M. Munsell, S. Plevritis, P. Ravdin, C. Schechter, B. Sigal, M. Stoto, N. Stout, N. van Ravesteyn, J. Venier, M. Zelen, and E. Feuer (2009), "Effects of Mammography Screening Under Different Screening Schedules: Model Estimates of Potential Benefits and Harms," *Annals of Internal Medicine*, 151, 738-747.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C. (2004a), "Social Learning from Private Experiences: The Dynamics of the Selection Problem," *Review of Economic Studies*, 71, 443-458.

Manski, C. (2004b), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221-246.

Manski, C. (2011), "Interpreting and Combining Heterogeneous Survey Forecasts," in M. Clements and D. Hendry (editors), *Oxford Handbook on Economic Forecasting*, Oxford: Oxford University Press, 457-472.

Manski, C. (2016), "Interpreting Point Predictions: Some Logical Issues," *Foundations and Trends in Accounting*, 10, 238-261.

Manski, C. (2018), "Reasonable Patient Care under Uncertainty," *Health Economics*, 27, 1397-1421.

Manski, C. (2019), *Patient Care under Uncertainty*, Princeton: Princeton University Press, forthcoming.

Manski, C. and J. Pepper (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997-1010.

Manski, C. and J. Pepper (2009), "More on Monotone Instrumental Variables," *The Econometrics Journal*, 12, S200-S216.

Manski, C. and J. Pepper (2013), "Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections," *Journal of Quantitative Criminology*, 29, 123-141.

Manski, C. and A. Tetenov (2016), "Sufficient Trial Size to Inform Clinical Practice," Proceedings of the *National Academy of Sciences*, 113, 10518-10523.

McNees, S. (1992), "The Uses and Abuses of 'Consensus' Forecasts," *Journal of Forecasting*, 11, 703-710.

National Center for Health Statistics (2019), *National Health and Nutrition Examination Survey*, https://www.cdc.gov/nchs/nhanes/index.htm, accessed January 13, 2019.

Spiegelhalter D., L. Freedman, and M. Parmar (1994), "Bayesian Approaches to Randomized Trials" (with discussion), *Journal of the Royal Statistics Society Series A*, 157, 357-416.

Spiegelhalter, D. (2004), "Incorporating Bayesian Ideas into Health-Care Evaluation," *Statistical Science*, 19, 156-174.

Stoye, J. (2009), "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70-81.

Stoye, J. (2012), "Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 138-156.

Stanley, T. and S. Jarrell (1989), "Meta-regression Analysis: a Quantitative Method of Literature Surveys," *Journal of Economic Surveys*, 3, 161-170.

Surowiecki, J. (2004), *The Wisdom of Crowds*, New York: Random House.

Tipton, E., J. Pustejovsky, and H. Ahmadi (2018), "A History of Meta-Regression: Technical, Conceptual, and Practical Developments between 1974 and 2018," Department of Statistics, Northwestern University.

Thompson, S. and J. Higgins (2002), "How Should Meta-Regression Analyses be Undertaken and Interpreted?" *Statistics in Medicine*, 21, 1559-1573.

Wald, A. (1950), *Statistical Decision Functions*, Wiley: New York.