

Toward Credible Patient-centered Meta-analysis

Charles F. Manski

Abstract: Meta-analysis is widely used to combine the findings of multiple disparate studies of health risks or treatment response. Meta-analysis often uses a random-effects model to express heterogeneity across studies. The model interprets a weighted average of study-specific estimates as an estimate of a mean parameter across a hypothetical population of studies. The relevance of this methodology to patient care is not evident. Clinicians need to assess risks and choose treatments for populations of patients, not for populations of studies. This article draws on econometric research on partial identification to propose principles for patient-centered meta-analysis. One specifies a patient prediction of concern and determines what each available study reveals. Given common imperfections in internal and external validity, studies typically yield credible set-valued rather than point predictions. Thus, a study may enable one to conclude that a probability of disease, or mean treatment response, lies within a range of possibilities. Patient-centered meta-analysis would combine the findings of multiple studies by computing the intersection of the set-valued predictions that they yield.

Keywords: Clinical decisions; Partial identification; Predicting treatment response; Research synthesis; Risk assessment

(*Epidemiology* 2020;31: 345–352)

Ideally, patient care should benefit from the many studies that analyze evidence on health risks and treatment response. In practice, difficulties arise when clinicians attempt to combine findings from multiple disparate studies. They must somehow interpret the mass of information provided by evidence-based research.

To inform clinicians, medical researchers sometimes perform systematic reviews of sets of studies. Systematic review is a subjective process, akin to exercise of clinical judgment. Seeking a more objective way to combine findings, statisticians have proposed meta-analysis; see, for example,

Editor's Note: A related commentary on this article appears on p. 353.

Submitted April 17, 2019; accepted February 6, 2020.

From the Department of Economics, Institute for Policy Research Northwestern University, Evanston, IL.

The author reports no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Charles F. Manski, Department of Economics, Northwestern University, 2211 Campus Drive, Evanston, IL 60208. E-mail: cfmansi@northwestern.edu.

Copyright © 2020 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 1044-3983/20/3103-0345

DOI: 10.1097/EDE.0000000000001178

Hedges and Olkin.¹ Application of meta-analysis has become common in the literature on evidence-based medicine.

The prevailing practice of meta-analysis is easy to motivate when combining findings poses a purely statistical problem. Suppose that one wants to estimate as precisely as possible a parameter characterizing a study population. 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 samples and compute the estimate using all the data. Often, however, the raw data are unavailable. Instead, multiple parameter estimates are available, each computed with data from a different sample. Meta-analysis proposes methods to combine these estimates. A well-motivated proposal in this setting is to compute a weighted average of estimates, the weights varying with sample size to minimize variance.

Meta-analysis as described above is uncontroversial, but its applicability is 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 designs. The studies may examine distinct populations whose members may have different probabilities of disease or different distributions of treatment response. The protocols for administration of treatments and measurement of outcomes may vary.

Meta-analysis often uses a random-effects model proposed by DerSimonian and Laird² to express heterogeneity across disparate studies. The model interprets a weighted average of study-specific estimates as an estimate of a mean parameter across a hypothetical population of studies. The relevance of this weighted-average estimate to patient care is not evident. Clinicians need to assess risks and choose treatments for populations of patients, not populations of studies. To express this distinction succinctly, I will say that clinicians should want meta-analysis to be *patient centered* rather than *study centered*.

This article proposes principles for patient-centered meta-analysis. I view the problem of combining study findings from the perspective of a clinician (or clinical team) treating a specific patient. The clinical perspective may differ from that of a medical researcher who performs a study-centered meta-analysis to summarize multiple studies in a journal article. The clinician observes some patient covariates, such as health history and the results of diagnostic tests. Conditioning on these covariates, the clinician wants to predict risk

of illness or a treatment outcome. This prediction problem is ubiquitous in clinical practice.

I consider how a clinician, behaving as an active interpreter of research rather than a passive reader of systematic reviews or study-centered meta-analyses, should interpret evidence that may be relevant to prediction. The clinician should recognize that both statistical imprecision and identification problems limit the informativeness of studies. 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 inferential difficulties that persist when sample size grows without bound.

Identification problems arise from imperfections in the internal and external validity of studies. Common imperfections in internal validity include missing data, measurement of surrogate outcomes, incomplete understanding of treatment selection in observational studies, and noncompliance in trials. Common imperfections in external validity are differences between the patients and treatments studied in the research and those under consideration in clinical practice.

The prevailing practice in medical research, using methods developed by biostatisticians and epidemiologists, has been to make assumptions strong enough to resolve identification problems, leaving statistical imprecision as the sole limit on the informativeness of studies. For example, when analyzing observational data, it is common to assume that treatment selection is random, conditional on specified patient covariates. When analyzing trial data, it is common to abstract from imperfections in external validity, mentioning them only in the discussion sections of articles. Thus, research articles report point-estimates for survival probabilities, average treatment effects, and other parameters of clinical relevance. Articles express statistical imprecision through confidence intervals or standard errors.

Coherent patient-centered meta-analysis can be performed by interpreting the evidence in each available study with assumptions strong enough to resolve identification problems. This done, each study may be used to form a consistent point estimate for a patient prediction of interest. Then, the only task of meta-analysis is to combine the estimates in a manner that improves statistical precision. In the absence of identification problems, forming a suitable weighted average of estimates is often a well-motivated way to improve precision.

Coherency of analysis, however, does not imply credibility. Beginning around 1990, econometric research on partial identification has argued that the assumptions required to resolve identification problems are typically too strong to be credible. Partial identification analysis has shown that research performed with credible assumptions does not generally reveal disease probabilities and treatment effects precisely, even as sample size grows without bound. However, research may yield credible bounds on these and other parameters. Classical

confidence intervals may be used to jointly quantify partial identification and statistical imprecision.

Several books^{3–5} provide in-depth expositions at different technical levels for students and researchers in economics and statistics. Two nontechnical monographs^{6,7} present basic ideas and applications for practitioners in public policy and patient care, respectively. There also are multiple review articles with different emphases.^{8–10}

Consideration of partial identification provides foundations for credible patient-centered meta-analysis. Suppose that multiple studies are available. Each enables one to credibly conclude that a parameter of interest, say a patient's probability of disease or mean treatment response, lies within some bound. It follows logically that the true value of the parameter lies in the intersection of these study-specific bounds.

Consider, for example, combining findings from an observational study and a randomized trial. Jointly considering internal and external validity, an observational study and a trial may each credibly 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 study of set-intersection to combine the findings of multiple studies has long been part of econometric research on partial identification, beginning in Manski¹¹ and subsequently generalized.^{4,12–15} To my knowledge, it has not, however, been specifically proposed as an alternative to traditional study-centered meta-analysis. This is the methodological contribution of the present article.

The study-centered meta-analysis is well known, so I provide here only a brief discussion that focuses on its characterization of heterogeneity across disparate studies. Although econometric research on partial identification has been in progress for the past 30 years, I am aware that this literature is not familiar to many researchers immersed in the study-centered meta-analysis. Indeed, it may initially appear foreign. I explain its application to patient-centered meta-analysis in two stages. I present basic concepts when one study is available and gives two illustrations. Both concern treatment for hypertension. When partial identification with one study is understood, extension to settings with multiple studies is conceptually simple. I explain, showing how set-intersection arises and giving examples.

EXPRESSING HETEROGENEITY IN STUDY-CENTERED META-ANALYSIS

Glass,¹⁶ who introduced the term *meta-analysis*, recognized the challenge of combining findings from disparate studies. He wrote (p. 358): “The tough intellectual work in many applied fields is to make incommensurables commensurable, in short, to compare apples and oranges.”

To operationalize comparison of “apples and oranges,” meta-analyses frequently view disparate studies through the lens of the random-effects model proposed by DerSimonian and Laird.² A random-effects model recognizes that each study

examined in a meta-analysis may concern a distinct value of a parameter of interest. It places structure on the heterogeneity of parameters across studies by supposing that each study is drawn at random “from a population of potential studies” (p. 181). It interprets a weighted average of the estimates across studies as an estimate of the mean parameter value across studies. It uses the variance of the parameters to measure the extent to which parameter values vary across studies.

Examples: Medical researchers have used random-effects models to perform numerous meta-analyses of studies evaluating treatments for many diseases. For example, Buchwald et al.¹⁷ combined the findings of 134 studies of the outcomes of bariatric surgery. The studies included five randomized trials, 28 nonrandomized but somehow otherwise controlled trials, and 101 “uncontrolled case series.” The studies were performed in different countries and followed patients for different periods of time. They measured weight loss, an outcome of interest, in multiple ways. To summarize findings, 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.

Chen and Parmigiani¹⁸ performed a meta-analysis of 10 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 was no consensus across the studies, which reported heterogeneous estimates with data from different study populations.

The mean and variance of parameter values across the studies examined in a meta-analysis are well-defined quantities if one assumes that each study is drawn at random from a population of possible studies. However, it is not obvious how to conceptualize “a population of possible studies,” nor why the available studies should be considered a random sample from this population, nor how the result is relevant to patient care. Considering sclerotherapy trials, Thompson¹⁹ questioned clearly the practice of averaging findings, stating (p. 1352):

Given the clinical heterogeneity, we do not know to which endoscopic technique, to which selection of patients, or in conjunction with what ancillary clinical management such a conclusion is supposed to refer. It is some sort of ‘average’ statement that is not easy to interpret quantitatively in relation to the benefits that might accrue from the use of a specific clinical protocol.

In a recent retrospective article, DerSimonian and Laird²⁰ acknowledge these concerns but belittle them, writing:

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. (p. 142)

The final sentence of the above statement refers to two objectives for meta-analysis, “providing an overall summary of what has been learned” and “a quantitative measure of how results differ, above and beyond sampling error.” In the random-effects model, the mean and variance of the parameter value across studies accomplish these objectives.

Some researchers schooled in meta-analysis have wanted to characterize heterogeneity across studies more deeply than through the variance of parameter values. To this end, they have used *metaregressions* to describe how findings vary with observed attributes of studies. Several articles^{21–23} provide perspectives. Metaregressions essentially perform study-centered meta-analysis within subpopulations of studies with specified attributes. They characterize how parameter values vary within and across subpopulations of studies. Metaregressions characterize how parameters vary across subpopulations of patients only in special cases where the studies being combined differ only in the composition of their patients.

Although averaging predictions across studies has been the norm in meta-analysis, investigators occasionally present a range of predictions and describe the range as measuring uncertainty. eAppendix 1; <http://links.lww.com/EDE/B647> uses breast-cancer risk assessment to illustrate.

PARTIAL IDENTIFICATION WITH EVIDENCE FROM ONE STUDY

Basic Ideas

Henceforth, I consider a clinician who wants to personalize patient care. One might think of personalized care as literally specific to an individual patient, but knowledge to support complete personalization is generally not available. Personalized care usually means care that varies with observed patient covariates.

A common clinical prediction objective is to probabilistically assess health risks or predict treatment response conditional on observed patient covariates. Existing studies may provide relevant information, but they may not yield fully accurate probabilistic predictions. The question is how a clinician may reasonably use the available findings.

In this section, I suppose that a single study has been performed. The study was conducted in some population of patients, measuring study-specific outcomes and patient

covariates. In studies of treatment response, the treatments administered may have been study specific as well. The study design may suffer from any or all the imperfections of internal and external validity mentioned in the Introduction.

Research on partial identification studies the conclusions about parameters of interest that hold when evidence from a study is combined with alternative assumptions. The generic result is a determination that a parameter of interest lies within a set of possibilities, called its identification region or identified set. The literature has characterized identification regions for various parameters, under alternative assumptions. The parameters analyzed include mean outcomes, quantiles, and spread parameters. The assumptions considered aim to be credible in observational studies or randomized trials.

A central theme has been to characterize how the conclusions drawn in empirical inference vary with the strength of the assumptions maintained. One may be tempted to maintain strong assumptions, to draw strong conclusions. However, there is a tension between the strength of assumptions and their credibility, described in Manski⁴ (p. 1) as “The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.” This implies that analysts face a dilemma as they decide what assumptions to maintain. Stronger assumptions yield conclusions that are more powerful but less credible.

The literature also emphasizes the distinction between refutable and nonrefutable assumptions. An assumption is not refutable if the identification region obtained with the evidence obtained in a specified manner is necessarily nonempty. Leading examples are assumptions on distributions of missing data. Missing data being unobserved, available evidence logically cannot refute any assumptions regarding the values of missing data. Assumptions are refutable if combining them with the evidence could conceivably imply an empty identification region. If this occurs, one should conclude that the assumption is incorrect.

Manski⁵ emphasizes that one should not confuse the refutability of an assumption with its credibility. Refutability is a matter of logic and credibility is a subjective matter. An assumption is refutable if it is inconsistent with some possible configuration of the empirical evidence and it is nonrefutable otherwise. Credibility is a property of an assumption and the person contemplating it. An assumption is credible to the degree that someone thinks it so.

To formalize concepts in the context of clinical prediction, let the patient under consideration be called patient “0” and have clinically observed covariates x_0 . Let y denote a patient outcome of clinical concern. A common concern is whether a patient will survive for a specified period; then $y = 1$ if the patient survives and $y = 0$ otherwise. Another common concern is length of survival; then y measures length of survival.

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 treatments. For each $t \in T$, let $y(t)$ denote the health outcome that the patient would experience with treatment t . The probabilistic prediction of treatment response means that the clinician wants to know the conditional distributions $P[y(t)|x_0]$, $t \in T$. These are distributions of outcomes among patients with covariates x_0 . They are not distributions across some conjectured population of studies.

Clinicians often want to learn mean outcomes. Then, the quantity of interest may be $E(y|x_0)$ or $E[y(t)|x_0]$, $t \in T$. The identification region for $E(y|x_0)$ given specified maintained assumption is some set E_0 on the real line and that for $E[y(t)|x_0]$, $t \in T$ is some set H_0 of dimension $|T|$. In many applications, the former set is an interval and the latter a hyperrectangle of dimension $|T|$.

To illustrate partial identification analysis, I next summarize research on two common identification problems stemming from missing data and use medical applications to illustrate. I then describe the formation of confidence regions that jointly characterize partial identification and statistical imprecision.

Partial Identification of Mean Treatment Response with an Observational Study

Manski¹¹ derived nonparametric bounds on a mean treatment response obtained from an observational study. The bounds suppose that the study examines data from the clinically relevant patient population, who are administered the treatments of interest. Thus, there are no imperfections of external validity. However, as is common in observational studies, the researcher/clinician lacks knowledge of the process of treatment selection.

Suppose that a clinician wants to learn $E[y(t)|x_0]$, t being a treatment of interest. Let z denote the treatment received by a member of the study population. Let $P(z = t|x_0)$ be the fraction of persons in the study population who receive t , among those with covariates x_0 . The Law of Iterated Expectations gives

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

where $E[y(t)|x_0, z = t]$ is mean treatment response within the group who have covariates x_0 and who receive treatment t . $E[y(t)|x_0, z \neq t]$ is mean response for those who receive another treatment. An observational study reveals $P(z = t|x_0)$ and $E[y(t)|x_0, z = t]$ as sample size increases. $E[y(t)|x_0, z \neq t]$ is counterfactual, hence unobservable.

An informative bound emerges if $y(t)$ has known bounded range, say $[y_L, y_U]$. Then, making no assumptions about treatment selection in the study, we can conclude that $E[y(t)|x_0]$ lies in the interval

$$\begin{aligned} E[y(t)|x_0, z = t] \times P(z = t|x_0) + y_L \times P(z \neq t|x_0) &\leq E[y(t)|x_0] \\ &\leq E[y(t)|x_0, z = t] \times P(z = t|x_0) + y_U \times P(z \neq t|x_0). \end{aligned} \quad (2)$$

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.

The bounds take a particularly simple form when $y(t)$ is a binary outcome taking the value 0 or 1, as is the case in survival analysis. Then, the objective is to learn the survival probability $E[y(t)|x_0] = P[y(t) = 1|x_0]$. The values of y_L and y_U logically are 0 and 1, respectively. The bound reduces to

$$\begin{aligned} & P[y(t) = 1|x_0, z = t] \times P(z = t|x_0) \leq P[y(t) = 1|x_0] \\ & \leq P[y(t) = 1|x_0, z = t] \times P(z = t|x_0) + P(z \neq t|x_0). \end{aligned} \quad (3)$$

When outcome $y(t)$ is not logically bounded, as with a binary outcome, clinical judgment is required to set the feasible range $[y_L, y_U]$ for counterfactual outcomes. In this regard, it is important to note that placing this bound on all counterfactual realizations of $y(t)$ is stronger than necessary. To obtain bound (2), it suffices to assume that the counterfactual mean outcome $E[y(t)|x_0, z \neq t]$ lies in the interval $[y_L, y_U]$.

Using NHANES Data to Bound Response to Treatment for Hypertension

The National Health and Nutrition Examination Survey (NHANES) Survey²⁴ is an ongoing observational study that assesses the health status of the American population through a continuous cross-sectional survey. Among many topics, the survey measures blood pressure and enquires about treatment for hypertension. 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. eAppendix 2; <http://links.lww.com/EDE/B647> provides further detail.

NHANES does not suffer from serious problems of external validity. It 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 imperfect internal validity in the absence of knowledge of the process of treatment selection in clinical practice.

eTable 1; <http://links.lww.com/EDE/B647> gives the findings on average systolic blood pressure (SBP) for the 4837 subsample respondents across the 2007–2016 period. 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 and would be 142.2 if none were treated. However, the NHANES being an observational study one may not find it credible to assume that treatment is random.

Computation of bound (2) makes no assumption about treatment selection, but it requires credible values for y_L and y_U . To illustrate, I use 100.7 and 182.0, which are the observed 2.5 and 97.5 percentile values of SBP among the 4837 NHANES subsample members; a clinician with different judgment could use other values. Computation of (2) with $y_L = 100.7$ and $y_U = 182.0$ yields these bounds on mean SBP with and without treatment: [133.4, 138.2] and [103.2, 179.6].

The fact that 0.94 of the persons diagnosed with hypertension receive treatment makes the NHANES problem of

internal validity highly asymmetric. 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]. The data yield the narrow bound [133.4, 138.2] on mean SBP if all older Americans with hypertension were treated. Thus, the NHANES data reveal little about patient outcomes without treatment, but the data are highly informative about outcomes with treatment.

Identification with Missing Data on Patient Outcomes or Covariates

Missing data on patient outcomes and covariates are a frequent occurrence in trials and observational studies. Researchers commonly assume that data are missing at random. This done, they often report findings only for sampled patients with complete data, discarding those with incomplete data. Or they impute missing values and report findings for all patients, acting as if imputations are actual data values. Either way, researchers report point estimates of treatment effects.

A sequence of partial identification analyses has sought to understand how missing data may affect research conclusions. It is useful to begin by asking what one can learn about treatment response without any knowledge of the process generating missing data. Conclusions drawn in this manner are weaker, but more credible than ones drawn by assuming that data are missing at random or by making another assumption.

Inference without assumptions about the nature of missing data basically is a matter of contemplating all possible configurations of the missing data. Doing so generates the identification region. The challenge is to characterize the identification region in a tractable way, so applied researchers can use the findings.

Manski^{11,25,26} shows that analysis is simple when only outcome data are missing. The most transparent case occurs when the objective is to learn the success probability for a treatment when the outcome of interest is binary (success or failure). Then, the smallest and largest possible values of the success probability are determined by conjecturing that all missing outcomes are failures or successes, respectively. The same reasoning holds when the outcome of interest is a patient's remaining life span and the objective is to learn the mean or median outcome that may occur. The smallest and largest possible values of the mean or median are determined by conjecturing that all patients with missing outcomes die immediately or live as long as humanly possible.

Analysis is more complex when some sample members have missing outcome data, some have missing covariate data, and some have jointly missing outcome and covariate data. Horowitz and Manski^{27,28} study these settings. The latter article provides an illustrative application to a trial of treatments for hypertension, described below.

Using a Trial with Missing Data to Bound Response to Treatment for Hypertension

Horowitz and Manski²⁸ analyzed identification of treatment response when a trial is performed, but some patient

outcome or covariate data are missing. Focusing on cases in which outcomes are binary (success or failure), the researchers derived sharp bounds on success probabilities without assumptions about the distribution of missing data. They applied the findings to a trial on treatments for hypertension.

Materson et al.²⁹ reported a randomized trial comparing seven treatments for hypertension. The measured outcome was binary, with $y = 1$ if a criterion for success in reducing blood pressure was met and $y = 0$ otherwise. The authors examined how treatment response varies with the race and age of the patient. There were no missing data on these covariates. The authors performed an intention-to-treat analysis that interpreted attrition from the trial as lack of success; from this perspective there were no missing outcome data either. See eAppendix 3; <http://links.lww.com/EDE/B647> for details.

Horowitz and Manski used the trial data to examine how treatment response varies with another covariate that does have missing data. This is the biochemical indicator *renin response*, taking the values (low, medium, and high). Renin response was measured at the time of randomization, but data were missing for some subjects in the trial. The new analysis also removed the intention-to-treat interpretation of attrition as lack of success. Instead, it viewed subjects who attrit as having missing outcome data. The pattern of missing covariate and outcome data is shown in eTable 2; <http://links.lww.com/EDE/B647>. eTable 3; <http://links.lww.com/EDE/B647> shows the estimated bounds on treatment success probabilities.

To focus on identification, it is best to ignore sampling imprecision and suppose that the estimates are population bounds rather than estimates of the bounds. Many bounds are sufficiently narrow to enable one to conclude that certain treatments are dominated—that is, surely inferior to others. Without imposing assumptions on the distribution of missing data, a clinician can reject treatments 1, 6, and 7 for all patients, reject treatment 3 for patients with medium renin response, and determine that treatment 5 is optimal for patients with low renin response.

Measuring Statistical Imprecision in Settings with Partial Identification

The above discussion focused on identification, abstracting from statistical imprecision. In practice, samples may be small enough that imprecision is a serious concern.

The statistics literature on point estimation of population parameters uses confidence sets to measure uncertainty created by sampling variation. Recall the standard definition of a confidence set for a real parameter θ . One specifies a *coverage probability* α , where $0 < \alpha < 1$. One considers alternative ways to use the sample data to form sets on the real line. Let $C(\cdot)$ be a set-valued function that maps the data into a set on the line. Thus, for each possible value ψ of the sample data, $C(\psi)$ is the set computed when the data are ψ . Then, $C(\cdot)$ gives an α -confidence set for θ if $\text{Prob}[\psi: \theta \in C(\psi)] = \alpha$. In words, an α -confidence set contains the true value of θ with probability

α as the sampling process is engaged repeatedly to draw independent data samples. It typically is not possible to determine the exact coverage probability of a confidence set. Hence, statisticians seek asymptotically valid confidence sets, whose coverage probabilities converge to α as the sample size grows.

Although the statistics literature has focused on parameters that are point-identified, the standard definition of a confidence set also applies to parameters that are partially identified. In addition, one can define confidence sets for identification regions. Let $H(\theta)$ denote the identification region for θ . Then $C(\cdot)$ gives an α -confidence set for $H(\theta)$ if $\text{Prob}[\psi: H(\theta) \subset C(\psi)] = \alpha$. An α -confidence set for $H(\theta)$ necessarily covers θ with probability at least α . This holds because the true value of θ lies in $H(\theta)$; hence, $\text{Prob}[\psi: \theta \in C(\psi)] \geq \text{Prob}[\psi: H(\theta) \subset C(\psi)]$. Imbens and Manski³⁰ demonstrate that when a parameter is partially identified, an α -confidence set for the parameter may be strictly smaller than the corresponding confidence set for the identification region.

A large econometric literature developing asymptotically valid confidence sets for partially identified parameters and their identification regions has developed over the past 20 years. An early example was the use of the bootstrap by Horowitz and Manski²⁸ to form confidence intervals for identification regions in the setting described earlier. Later articles^{9,10} review subsequent theoretical research.

The literature has also developed hypothesis tests, much as in the classical statistical literature on hypothesis testing. For example, consider a simple null hypothesis that a parameter of interest has a specific value. Suppose that one uses the sample data to estimate the identification region for the parameter and computes the distance from the estimated region to the value specified under the null hypothesis. If this distance is suitably large, it is intuitive to reject the null hypothesis. The literature develops formal tests of this type, with the traditional objectives of controlling size and then maximizing power.⁹

PARTIAL IDENTIFICATION WITH EVIDENCE FROM MULTIPLE STUDIES

Basic Ideas

The previous section considered one study in isolation. Meta-analysis aims to aggregate findings across multiple studies. Suppose that K studies have been performed, each partially identifying a parameter of interest. Then, a potential value of the parameter is consistent with everything learned in the K studies if and only if it lies within each of the K identification regions. Thus, the identification region combining the studies is the intersection of the regions obtained with each study. This is the simple logic of set intersection.

Set intersection is easy to perform when the parameter is real valued and the K identification regions 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.

Example: Let three studies of patient survival be available. Suppose that their identification regions for the probability of survival are [0.4, 0.7], [0.2, 0.6], and [0.5, 1]. Set intersection yields [0.5, 0.6] as the bound on survival probability obtained by combining the studies.

This numeric example is interesting because it demonstrates a subtlety in the operation of set intersection. One might intuit that when K studies yield identification regions of different sizes, the studies yielding the smallest identification regions would play the most prominent role when combining findings. The example shows that this need not be the case. The lengths of three study-specific intervals are 0.3, 0.4, and 0.5, respectively. Yet, the set intersection is determined entirely by the second and third intervals. The lesson is that the result of set intersection depends on the joint positioning of the sets, not only by their sizes.

When considering statistical imprecision, combining studies that partially identify a parameter does not raise conceptual issues beyond those discussed earlier. Thus, one may seek to form a confidence set that covers either the parameter or the identification region with specified probability. And one may use sample data to test hypotheses. Although testing and formation of confidence sets raise no new conceptual issues, the subtlety of set intersection does make it challenging to develop tests and confidence sets that are tractable to compute and have desired statistical properties. Kreider and Pepper³¹ and Chernozhukov et al.¹⁵ study these matters. The former article applies simple heuristics and the latter develops formal asymptotic statistical theory.

Intersecting identification regions for a parameter of clinical interest differs markedly from use of random-effects models in study-centered meta-analysis. Medical applications of random-effects models make assumptions strong enough to point-identify the parameter value that pertains to each study. They then focus attention on the mean parameter value across studies.

Partial Identification with K Observational Studies

To illustrate set intersection, reconsider the problem of identification with an observational study examined earlier, except that now one observes evidence from K observational studies rather than one study. Manski¹¹ observed that $E[y(t)|x_0]$ lies in interval (2) computed for every $k \in K$. Hence, it lies within the intersection of the $|K|$ study-specific intervals. Subscripting observable study-specific quantities by the appropriate label k, the result is the intersection interval

$$\max \{E_k[y(t)|x_0, z_k = t] \cdot P_k(z_k = t|x_0) + y_L \cdot P(z_k \neq t|x_0)\} \leq E[y(t)|x_0] \quad (4)$$

$$k \in K$$

$$\leq \min \{E_k[y(t)|x_0, z_k = t] \cdot P_k(z_k = t|x_0) + y_U \cdot P_k(z_k \neq t|x_0)\}.$$

$$k \in K$$

Although set intersection is simple when considering mean outcomes, it is more complex when considering

complete distributions. Some results on the latter question are developed in Manski.^{4,32} Manski⁴ characterizes abstract set intersection when multiple studies use different sampling processes to examine different sub-populations of a population of interest. Manski³² combines findings on treatment response obtained from observational studies of multiple successive cohorts of persons. Observation of more cohorts enables one to intersect more sets, thereby increasing knowledge of the distribution of treatment response.

Partial Identification Using Evidence from Trials with Noncompliance

Another illustration of set intersection arises in analysis of treatment response using evidence from randomized trials with noncompliance. In general, a trial with K treatment arms can be conceptualized as research in which K observational studies have been performed and the outcomes observed. The study k yields evidence on the outcomes of subjects who are randomly assigned to treatment k.

In a trial with perfect compliance, the intersection interval (4) holds with the special property that $P_k(z_k = k|x_0) = 1$ and $P_k(z_k = j|x_0) = 0$ for all treatments $j \neq k$. Hence, for each treatment k, (4) reduces to the equality $E[y(k)|x_0] = E_k[y(k)|x_0, z_k = k]$.

In a trial with imperfect compliance, (4) still holds but the equalities $P_k(z_k = k|x_0) = 1$ and $j \neq k$ do not hold. Robins³³ recognized this and reported (4) as a bound on $E[y(k)|x_0]$. Later, Balke and Pearl³⁴ examined the noncompliance problem afresh and recognized that random assignment of treatments implies restrictions not only on mean treatment response but on the joint distribution of treatment response; that is, on $P[y(k), k = 1, \dots, K | x_0]$. Considering the special case with a binary outcome and two treatments, they showed that the identification region for mean treatment response using the full force of randomized assignment is sometimes smaller than the intervals given in (4). This is a subtle finding, subsequently generalized by Kitagawa.¹⁴

Partial Identification Combining Observational Studies and Trials

Combining findings from observational studies and trials has long been a concern in research on evidence-based medicine. The influential Cochrane Handbook³⁵ views trials as qualitatively superior to observational studies. The Handbook discusses the GRADE approach to rating the certainty of a body of evidence, the four rating levels being high, moderate, low, and very low. GRADE recommends that the “high” rating should be reserved for evidence from certain randomized trials. The Cochrane authors write (section 12.2.1): “Review authors will generally grade evidence from sound observational studies as low quality.”

Developers of clinical practice guideline often act accordingly, valuing trial evidence more than observational studies. Indeed, guideline developers sometimes use only trial evidence, excluding observational studies from consideration.

An example is James et al.³⁶, which reports new guidelines for management of hypertension. The authors write (p. 508): “The panel limited its evidence review to RCTs because they are less subject to bias than other study designs and represent the gold standard for determining efficacy and effectiveness.”

From the perspective of patient-centered meta-analysis, it is misguided to *prima facie* favor trials over observational studies. Jointly considering internal and external validity, observational studies and trials may each credibly 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. eAppendix 4; <http://links.lww.com/EDE/B647> uses treatment of hypertension to illustrate.

DISCUSSION

A common 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 patient care is not evident. Patient-centered research should aim to inform risk assessment and treatment for populations of patients, not populations of studies.

This article has laid out principles for patient-centered meta-analysis. To cope with the identification problems that regularly afflict medical research, partial identification analysis, with computation of the intersection of set-valued predictions, can replace computation of weighted averages of point estimates.

ACKNOWLEDGMENTS

I am grateful to Michael Gmeiner for research assistance and comments. I have benefitted from the comments of an editor and several reviewers. A previous version of this paper was circulated under the title “Meta-analysis for Medical Decisions.”

REFERENCES

- Hedges L, Olkin I. *Statistical Methods for Meta-Analysis*. New York: Academic Press; 1985.
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177–188.
- Manski C. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press; 1995.
- Manski C. *Partial Identification of Probability Distributions*. New York: Springer-Verlag; 2003.
- Manski C. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press; 2007.
- Manski C. *Public Policy in an Uncertain World*. Cambridge, MA: Harvard University Press; 2013.
- Manski C. *Patient Care under Uncertainty*. Princeton: Princeton University Press; 2019.
- Tamer E. Partial identification in econometrics. *Ann Rev Economics*. 2010;2:167–195.
- Canay I, Shaikh A. Practical and theoretical advances in inference for partially identified models. In: Honoré B, Pakes A, Piazzesi M, Samuelson L, eds. *Advances in Economics and Econometrics: Eleventh World Congress*. Cambridge: Cambridge University Press; 2017:271–306.
- Molinari F. Econometrics with partial identification. In: Durlauf S, Hansen L, Heckman J, Matzkin R, eds. *Handbook of Econometrics*. Amsterdam: North Holland; forthcoming; 2019.
- Manski C. Nonparametric bounds on treatment effects. *Am Economic Rev Papers Proc*. 1990;80:319–323.
- Manski C, Pepper J. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*. 2000;68:997–1010.
- Manski C, Pepper J. How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *Rev Econ Stat*. 2018;100:232–244.
- Kitagawa T. Identification Region of the Potential Outcome Distributions under Instrument Independence, CeMMAP working paper CWP30/09. Available at: <https://www.cemmap.ac.uk/wps/cwp3009.pdf>. Accessed February 14 2020.
- Chernozhukov V, Lee S, Rosen A. Intersection bounds: estimation and inference. *Econometrica*. 2013;81:667–737.
- Glass G. Integrating findings: the meta-analysis of research. *Rev Res Educ*. 1977;5:351–379.
- Buchwald H, Avidor Y, Braunwald E, et al. Bariatric surgery: a systematic review and meta-analysis. *JAMA*. 2004;292:1724–1737.
- Chen S, Parmigiani G. Meta-analysis of *BRCA1* and *BRCA2* penetrance. *J Clin Oncol*. 2007;25:1329–1333.
- Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*. 1994;309:1351–1355.
- DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemp Clin Trials*. 2015;45(Pt A):139–145.
- Stanley T, Jarrell S. Meta-regression analysis: a quantitative method of literature surveys. *J Economic Surv*. 1989;3:161–170.
- Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21:1559–1573.
- Tipton E, Pustejovsky JE, Ahmadi H. A history of meta-regression: technical, conceptual, and practical developments between 1974 and 2018. *Res Synth Methods*. 2019;10:161–179.
- National Center for Health Statistics. *National Health and Nutrition Examination Survey*. 2019. Available at: <https://www.cdc.gov/nchs/nhanes/index.htm>. Accessed 13 January 2019.
- Manski C. Anatomy of the selection problem. *J Human Resources*. 1989;24:343–360.
- Manski C. The selection problem. In: Sims C, ed. *Advances in Econometrics, Sixth World Congress*. Cambridge, UK: Cambridge University Press; 1994:143–170.
- Horowitz J, Manski C. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *J Econometrics*. 1998;84:37–58.
- Horowitz J, Manski C. Nonparametric analysis of randomized experiments with missing attribute and outcome data. *J Am Statis Assoc*. 2000;95:77–84.
- Materson B, Reda D, Cushman W, et al. Single-drug therapy for hypertension in men: a comparison of six antihypertensive agents with placebo. *New Engl JMed*. 1993;328:914–921.
- Imbens G, Manski C. Confidence intervals for partially identified parameters. *Econometrica*. 2004;72:1845–1857.
- Kreider B, Pepper J. Disability and employment: reevaluating the evidence in light of reporting errors. *J Am Statis Assoc*. 2007;102:432–441.
- Manski C. Social learning from private experiences: the dynamics of the selection problem. *Rev Economic Studies*. 2004;71:443–458.
- Robins J. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Service Research Methodology: A Focus on AIDS*. Washington, DC: NCHSR, U.S. Public Health Service; 1989.
- Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. *J Am Statis Assoc*. 1997;92:1171–1176.
- Higgins Jpt, Thomas J, Chandler J, et al. (eds). *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Chichester (UK): John Wiley & Sons, 2019.
- James P, Oparil S, Carter B, et al. 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). *JAMA*. 2014;311:507–520.