Incorporating Both Randomized and Observational Data into a Single Analysis

Eloise E. Kaizar

Department of Statistics, The Ohio State University, Columbus, Ohio 43210; email: kaizar.1@osu.edu

Annu. Rev. Stat. Appl. 2015. 2:49–72

First published online as a Review in Advance on March 12, 2015

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-010814-020249

Keywords

nonrandomized, selection bias, generalizability, bias model, research synthesis, cross design synthesis

Abstract

Although both randomized and nonrandomized study data relevant to a question of treatment efficacy are often available and separately analyzed, these data are rarely formally combined in a single analysis. One possible reason for this is the apparent or feared disagreement of effect estimates across designs, which can be attributed both to differences in estimand definition and to analyses that may produce biased estimators. This article reviews specific models and general frameworks that aim to harmonize analyses from the two designs and combine them via a single analysis that ideally exploits the relative strengths of each design. The development of such methods is still in its infancy, and examples of applications with joint analyses are rare. This area would greatly benefit from more attention from researchers in statistical methods and applications.

1. INTRODUCTION

Much of modern medical and social research is geared toward testing or estimating a causal association between an agent or treatment and a particular outcome. Developers of causal inference methods, however, disagree on many topics, including even the definition of an effect. In recent years, these arguments have focused particularly on the ideal study design. Experimentalists emphasize the potential for strong internal validity afforded by randomization. That is, because randomization creates two groups that are on average balanced on pre-treatment characteristics, one may on average attribute post-intervention differences between the two groups to differences in treatment. This feature has led to medical research communities considering the randomized controlled trial (RCT) to be the gold standard in providing causal evidence. The added support of the evidence-based medicine movement has solidified the RCT as the ideal design for medical inference. For example, randomized trials are routinely placed near the top of so-called evidence pyramids (see, e.g., Howick et al. 2011) and given the highest quality in rating schemes (Moher et al. 1995). Social scientists have more recently begun to favor randomization in their work as well. (see, e.g., Cook et al. 2008, Green & John 2010).

Despite the potential internal validity asset of randomization, studies without randomization have some important advantages (e.g., Black 1996, Williams & Garner 2002, Cohen et al. 2004, Cook et al. 2008). For example, observationalists note that compared with randomized studies, nonrandomized studies can be more feasible; simpler to conduct; and inclusive of more diverse subjects, treatments, and outcomes. Researchers in many fields have been increasingly interested in the usefulness of observational designs, as evidenced by a multiplicity of recent meetings and conferences that highlighted the relative merits of nonrandomized designs (e.g., IOM 2013, Reeves et al. 2013). The continuing development of more appropriate methods for extracting evidence on causal effects from nonrandomized data has also helped to elevate the value of these designs in the minds of many researchers.

Although unequal valuation of different assets leads individual researchers to favor different designs, both randomized and nonrandomized studies are currently and will continue to be relevant for inquiries into a wide range of treatment effects. Growing recognition of the relative advantages of and comfort with modern methods for both randomized and nonrandomized studies creates a pressing need for methods to jointly analyze both types of data. A case study of the safety of pediatric antidepressant use (see sidebar, Case Study: Pediatric Antidepressant Safety) demonstrates the relative merits of an array of relevant data and the importance of basing decisions on all relevant evidence.

Joint analysis of multiple data sets is the subject of a rich literature and the focus of several organizations (e.g., the Cochrane Collaboration for medical research and the Campbell Collaboration for social science). Most work in this area relates to combining summaries from collections of studies with the same design and related to the same or similar questions in order to estimate a single treatment effect. Further, the emphasis has been on combining only high-quality studies, implying that only randomized studies are used. Although both the Cochrane and Campbell Collaborations allow the inclusion of nonrandomized data in research reviews, the lack of supporting tools for doing so belies the smaller value that has until recently been afforded these designs. For example, the Cochrane Collaboration is just now developing a tool for evaluating the quality of nonrandomized studies of interventions (Reeves et al. 2013). Pressures such as the need for policy makers to have timely information on relevant effects (Reeves et al. 2013) have pushed the field into recognizing the value of casting a broader net, and many researchers suggest that randomized and nonrandomized studies can provide complementary information (Grootendorst et al. 2010, Peinemann et al. 2013, Shrier et al. 2007).

CASE STUDY: PEDIATRIC ANTIDEPRESSANT SAFETY

Although psychiatric professionals had suspected an increase in treatment-induced suicides since the early 1990s (Teicher et al. 1990), intense study of this issue did not begin until a 2003 GlaxoSmithKline report to the US Food and Drug Administration (FDA) indicated some evidence for this causal relationship in its pediatric RCTs of Paxil (Hamad et al. 2006). Shortly thereafter, the FDA and its British counterpart [the Medicines and Healthcare Products Regulatory Agency (MHRA)] issued public health advisories warning of the increased risk. These advisories prompted a number of conflicting studies based on a wide array of data sources. In this case study, I highlight analyses that address issues of internal and external validity. [Bridge & Axelson (2008) provide a more comprehensive review of relevant studies.]

In an analysis typical of evidence-based medicine, the follow-up study conducted by the FDA relied exclusively on RCTs. Their meta-analysis of 24 trials found a roughly twofold increase in the risk of suicidal thoughts and behaviors (termed "suicidality") due to new-generation antidepressant drugs (Hamad et al. 2006), and, in 2006, the FDA strengthened its advisory to a black box warning. Although randomization may protect against selection bias, this so-called definitive study has several shortcomings, such as use of exclusion criteria (including baseline suicidality), reliance on a surrogate endpoint, and short study duration (Kaizar et al. 2006). Some of these limitations to the study of pediatric antidepressant safety using RCTs can be addressed via many types of alternative data sources, as the following examples demonstrate.

1. The role of exclusion criteria (or, more broadly, population sampling) can be examined via both variability among *RCTs* with different enrollment strategies (see, e.g., Kaizar et al. 2006) and comparisons with more representative samples. Greenhouse et al. (2008) took the latter approach to judging the generalizability of the FDA's collection of RCTs. Using a *nationally representative sample survey* from the United States (the Youth Risk Behavior Survey), they found the rate of suicidality events in the RCTs to be half that observed in the adolescent population at large. Such a substantial difference casts doubt on the direct relevance of the FDA's results for the treatment of suicidal depressed youth.
2. The large sample sizes and longer follow-up times of *insurance claims databases* allow researchers to use attempted and completed suicide as an outcome, rather than the more frequent surrogate outcome, "suicidal thoughts and behaviors." (There were no completed suicides in the 24 trials.) Analyses based on a private insurance claims database (Valuck et al. 2004) and a Medicaid claims database (Cooper et al. 2014) did not find any significant association between antidepressant use and suicide attempt. Note that regardless of study size and design, classification of observed actions according to the child's intent is difficult at best and a limitation of any study of suicide.
3. *National mortality statistics* allow us to focus on completed suicide only, as adjudicated by coroners or similar professionals. Gibbons et al. (2006) used the natural variation in antidepressant prescription rates (as estimated from a *pharmacy database*) among US counties to show that counties with higher rates of antidepressant use also tend to have lower rates of adolescent suicide, although the ecological data do not allow a direct causal conclusion to be made.
4. Because of the regulatory actions taken by the FDA, variation in suicide rates across time also implies that antidepressant use is protective for completed suicide. Studies based on *national mortality statistics* (Gibbons et al. 2007) and a large *longitudinal insurance claims database* (Lu et al. 2014) showed increases in US suicide rates after the FDA actions. Although this trend might be attributable to other changes that occurred during that same time period, these data add to the preponderance of nonrandomized evidence against increased risks of completed suicide due to antidepressant use.

Although these examples focus on US-based data, many of these same studies have been repeated in Europe, with similar conclusions.

CASE STUDY: PEDIATRIC ANTIDEPRESSANT SAFETY (CONTINUED)

This case study clearly shows that although regulatory agencies considered randomized data to be definitive, there is a wealth of nonrandomized data that provide evidence on many aspects of the underlying question of the safety of antidepressant drugs. Consideration of all of the available evidence may have led to a decision more focused on nuanced trade-offs related to the real, long-term risk of suicide.

As of now, there has been only cautious exploration into how to jointly analyze evidence from both randomized and nonrandomized studies. For example, a search of the medical literature in 2012 produced only 42 papers on the joint use of multiple study designs, whereas 2,270 meta-analyses were indexed in PubMed between 1995 and 2012 (Peinemann et al. 2013, Ioannidis et al. 2013). Given the potential for complementary analyses, explicit methods for and successful examples of syntheses are in very short supply.

This article aims to review the literature directly related to methods for incorporating evidence from randomized and nonrandomized designs in a single analysis, as well as some relevant supporting work. Both broad design categories contain many subclassifications. However, the key concepts discussed in this paper typically apply regardless of this detailed taxonomy. In particular, the term "observational study" is sometimes reserved to describe a particular subset of nonrandomized designs that excludes interventional studies such as uncontrolled or historically controlled trials. Instead, I use the terms "nonrandomized" and "observational" interchangeably, and, in particular, I identify specific designs for which this distinction may be important. In addition, a very large and hotly discussed literature is devoted to correcting for selection bias in nonrandomized studies. Although these corrections are important to syntheses across designs, a review of these methods is beyond the scope of this article.

The review is organized as follows. Section 2 discusses and carefully defines the causal effect that is often the estimand of interest. Section 3 reviews methods that estimate single effect sizes via linear combinations of study-specific effect sizes, which are direct extensions of fixed- and random-effects meta-analyses. In this section, I also review an approach that has recently been losing favor, which is to adjust the weight assigned to low-quality study effect estimates that may be biased. Section 4 contains more general approaches to adjusting each study for possible design shortcomings and combining studies via methods that acknowledge design differences, including the cross design synthesis method proposed by Eddy et al. (1992) and other parametric bias models. I place a particular focus on methods to estimate how generalizable the results of a study may be, as well as methods to construct more generalizable effect estimates. The article concludes with a summary of the existing literature, comments on possible future directions, and a plea for more methods development and examples of applications in this important area.

2. TREATMENT EFFECT DEFINITIONS

In this review, I consider only those analyses for which the goal is estimation of the effect of an intervention (as opposed to the effect of some factor that cannot be changed directly, such as a person's race). Even with such focus, there is no single uniform definition of a treatment effect. Definitional differences are extremely important to the relevance of various study designs and analysis choices. Of the many possible ways to specify a treatment effect, the individual treatment effect (ITE) is perhaps the most universally applicable, as knowledge of this effect for all individuals would permit perfect decision making (see sidebar, Individual Treatment Effect). Because we can

INDIVIDUAL TREATMENT EFFECT

The effect that an intervention has on a single individual, typically taken to be the difference between the outcome that would have been observed had the individual received the intervention and the outcome that would have been observed had the individual not received the intervention, is known as the individual treatment effect (ITE). [This definition of treatment effect can be extended to consider probability distributions for the outcomes of treatment versus lack of treatment, as well to cases in which multiple or continuous treatments are considered (see, e.g., Imai & van Dyk 2004).]

observe or have direct evidence on only one individual counterfactual outcome, we must rely on an indirect or modeling approach to estimate each ITE. For example, in a study in which randomization balances all covariates across treatment groups, the average outcome among those who received the active treatment may be a reasonable estimate for the average counterfactual outcome for those who received the control treatment, and vice versa. Thus, the difference between these two sample averages, the primary statistic in many randomized trials, is an estimator of the sample average treatment effect (SATE). Although not theoretically required, methods for causal inference often focus on average treatment effects, regardless of study design or statistical tools used. (One notable exception is in economics, in which distribution quantiles are typically of more interest than means.)

Counterfactual outcome: The outcome for one individual that would be observed under one of the possible interventions, regardless of the intervention that was actually received

Sample average treatment effect (SATE): The average ITE, averaged across the study participants

Imai et al. (2008) provide an excellent in-depth review of average effects that may be of interest, as well as elements of study design and analysis that can be used to improve estimates of these means. In particular, they recognize that estimating average causal effects involves two important and interrelated considerations: The first (Consideration 1) is how to estimate the (average) counterfactual outcomes (i.e., internal validity), and the second (Consideration 2) is how to define the population over which the treatment effect should be averaged (i.e., external validity). In the latter, population can be very broadly defined; for example, it can be defined to include not only individual demographics but also environment. The clear distinction between these two considerations is essential for joint analyses of randomized and nonrandomized data, as each design relates to these

POST-ASSIGNMENT SELECTION BIAS

In theory, simple averages are attractive estimators for counterfactual average outcomes for randomized trials. In practice, however, one must utilize them with caution, as the course of events occurring after assignment to treatment (and perhaps due to this assignment) but prior to outcome measurement can severely bias these estimators. For example, randomized subjects may drop out of a study for many important reasons, such as perception of an ineffective control treatment (likely inducing a positive bias in the estimate of the counterfactual average outcome for the control treatment) or adverse side effects of the active treatment (likely inducing a negative bias in the estimate of the counterfactual average outcome for the active treatment). Model-based approaches to missing data can correct these biases in some cases. As a second example, consider subjects who do continue to participate in the study but do not comply with the assigned treatment. One may be able to overcome this problem by using principle stratification to define latent groups based on potential compliance with assigned treatment and by using observed compliance to create unbiased estimates of the average treatment effect on the compliers instead of on the entire sample (Frangakis & Rubin 2002).

considerations very differently. Typically, trialists address Consideration 1 through randomization and dismiss Consideration 2 by arguing that the purpose of the study is to test for a beneficial treatment effect for any population subset (rather than to estimate the average magnitude of the effect in a well-defined population). Observationalists address Consideration 1 via statistical methods to adjust for selection bias (e.g., propensity score matching or instrumental variables) and rely on elements of study design, such as probability sampling, to define the population in Consideration 2.

Selection bias: Any bias in a causal estimator due to systematic imbalance between the treatment groups

Generalizability bias: Any bias in a causal estimator due to systematic differences between a study and the application of interest, including the population, treatment, and outcome

Hawthorne effect: Changes in the counterfactual outcome of an individual that result from the attention garnered by study participation

Mismatches between the treatment effect (as defined by the two considerations) and the study protocol may bias effect estimates. Errors in addressing Consideration 1 are termed "selection biases" (which could result from issues such as dropout, missing data, or errors in measurement), and errors with respect to Consideration 2 are called "generalizability biases." In addition to population considerations, the generalizability of an estimator encompasses other factors regarding the relationship between (*a*) interventions and outcomes observed within the confines of a study and (*b*) those that would be observed naturally. For example, critics of randomized trials in health care note that the definitions of treatment (e.g., counseling intervention provided by an expert) and outcome of interest (e.g., short-term follow-up or surrogate endpoint) often differ from those observed naturally once an intervention is adopted (e.g., novice counselor or long-term outcome) (Rothwell 2005). Finally, one must consider phenomena such as the Hawthorne effect, in which the act of conducting a study influences the outcomes observed in that study. For example, many segments of the population spurn participation in organized studies, thus skewing or truncating the population represented in any study requiring recruitment. When jointly considering randomized and nonrandomized studies, we must pay particular attention to the danger of influencing recruitment and retention by the very act of randomization (Heckman & Smith 1995).

Discrepancies in target treatment effect definitions are critical, as they may explain ostensibly different effect estimates derived from randomized and nonrandomized studies. To demonstrate this point, Hernán et al. (2008) compared the risk of coronary heart disease due to hormone replacement therapy as estimated in the Women's Health Initiative (WHI) RCT with that reported in the Nurses' Health Study (NHS) observational study. The original observational data analyses indicated that hormone users are at reduced risk of heart disease, whereas the randomized study indicated the reverse: Hormone users are at increased risk. Hernán et al. (2008) reanalyzed the NHS data mimicking the same design and analysis choices of the randomized WHI study, notably using regression adjustments to address possible selection bias and using similar inclusion criteria to address generalizability bias. (The reanalysis also carefully harmonized the definition of treatment by looking at only hormone initialization, not current use, in order to parallel the "intent to treat" analysis in the RCT.) This case study found that "much of the apparent WHI–NHS difference disappeared" (p. 773) once the definition of the treatment effect was standardized. Although many comparisons do find differences between effects estimated from studies using different designs (Peinemann et al. 2013), MacLehose et al. (2000) and Shadish et al. (2008) generally confirmed the results of Hernán et al. (2008), finding only small differences in effect estimates in comparisons between randomized and observational studies with comparable populations and adequate control for selection.

3. LINEAR COMBINATIONS OF STUDY-SPECIFIC EFFECTS

Regardless of design, when multiple studies provide evidence regarding the same or similar treatment effects, it is natural to consider combining these studies in a single analysis. As in the single-design case, analysts often do not have access to original study data and must rely on reported study-specific estimates to conduct their analysis. Thus, researchers also naturally consider extending single-design meta-analytic methods to accommodate studies with differing designs.

3.1. Fixed-Effects Meta-Analysis

Historically, meta-analyses have typically focused on combining study-specific effect estimates from published studies via a weighted average. For example, let $\{\hat{\theta}_j\}_{j=1}^J$ and $\{\hat{\sigma}_j\}_{j=1}^J$ represent estimators of the effect sizes and their corresponding standard errors from J separate studies. For randomized studies, the estimator is typically the difference in means or a log odds ratio comparing the active treatment and control arms. For nonrandomized studies, the estimator may be adjusted to eliminate selection bias (as mentioned in Section 2).

These meta-analytical estimators can be developed under two paradigms. The first supposes that each study estimate is unbiased for the true treatment effect. In this case, the most simplistic estimator is the simple average. Weighting each study-specific estimate according to its precision reduces the variance of the overall estimator, and the result is termed the fixed-effects estimator:

$$\hat{\theta}_{\mathrm{FE}} = \frac{1}{\sum_j 1/\hat{\sigma}_j^2} \sum_j \frac{\hat{\theta}_j}{\hat{\sigma}_j^2}.$$

Heuristically, the fixed-effects estimate gives more weight to studies with smaller variability (e.g., those with larger sample size) than to those with larger variability. When the study-specific estimates have approximately normal probability distributions, $\hat{\theta}_{\mathrm{FE}}$ is the maximum likelihood estimator for the linear model:

$$\hat{\theta}_j = \theta + \epsilon_j,$$

where ϵ_j have independent $N(0, \hat{\sigma}_i^2)$ distributions. Fixed-effects estimators can also be formulated within a generalized linear model framework, such as the following model for a continuous outcome:

$$\overline{Y_{ij}} = \mu_j + \theta T_{ij} + \epsilon_{ij},$$

where T_{ij} is an indicator of the active treatment group (equal to 1 for $i = 1$ and to 0 for $i = 2$) and ϵ_{ij} have independent $N(0, \hat{\sigma}_j^2)$ distributions.

Although this type of model is typically applied to collections of similarly designed studies, it can incorporate both randomized and nonrandomized data. Begg & Pilote (1991) used a fixed-effects model to jointly analyze a collection of randomized trials and separate uncontrolled trials of each treatment to estimate the comparative effect of allogeneic bone-marrow transplantation versus conventional chemotherapy in treating leukemia. Under the assumption that $\{\mu_j\}_{j=1}^J$ are independently and identically distributed (i.i.d.) with mean μ and finite variance τ^2, the uncontrolled trials contribute to the information about the treatment effect θ. [Begg & Pilote (1991) require normal distributions, but Li & Begg (1994) do not.] They note that if the mean outcome μ_j varies a lot across studies, this approach naturally discounts the contribution of the uncontrolled studies (Begg & Pilote 1991). In fact, for infinite τ^2, the uncontrolled studies do not contribute to the estimate of θ at all. For τ^2 close to 0, however, inference about the treatment effect θ can be greatly improved by synthesizing the uncontrolled studies along with the randomized trials. Estimation for both models is accomplished via maximum likelihood or empirical Bayes methods.

This model unrealistically specifies identical means for both randomized and nonrandomized studies. Begg & Pilote (1991) extended the model to include treatment-specific bias terms, η and ξ, for the uncontrolled trials:

$$\overline{Y_{ij}} = \mu_j + \theta T_{ij} + \eta S_j + \xi T_{ij} S_j + \epsilon_{ij},$$

where S_j is an indicator for an uncontrolled study. In this way, the estimand θ represents the gold standard treatment effect with strong internal validity. Unfortunately, because this model includes an interaction between treatment and study type, the uncontrolled trials now contribute only via

improved estimation of the between-study variance τ^2. Note that this model assumes constant bias (although some heterogeneity may be reflected in the variance estimate σ_{ij}^2) for all uncontrolled studies, a strong a priori assumption.

Consideration of bias for the uncontrolled studies points to the possibility that the true treatment effect is not uniform across studies. Even if randomization balances the treatment groups, it is seldom tenable to assume a constant treatment effect, even among RCTs, because the studies to be synthesized rarely have uniform protocols (e.g., they may have different participant selection, treatment application, environment, outcome measurement, and follow-up time). The second synthesis paradigm assumes such between-study heterogeneity of treatment effects. If the studies are thought to be randomly selected to be naturally representative of the real-world heterogeneity among populations, treatments, or other study features, then the simple average described above is still a reasonable estimate of the true average treatment effect (Peto 1987). Most researchers, however, turn to random-effects models instead.

3.2. Random-Effects Meta-Analysis

Random-effects models place a probability distribution on the study-specific effect sizes, $\{\theta_j\}_{j=1}^J$. Though not necessary, i.i.d. normality is typical:

$$\hat{\theta}_j \sim \mathrm{N}(\theta_j, \hat{\sigma}_j^2),$$
$$\theta_j \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \tau^2).$$

The usual target of inference, θ, is the average of the true study-specific effects, but it is interpreted as the "true" treatment effect. The numerous methods for its estimation include method-of-moments-based weighted averages, generalized linear models, and Bayesian hierarchical models (Amatya et al. 2014).

As with fixed-effects models, nothing about the random-effects model dictates study type. One can assume that randomized and nonrandomized studies have identical means and directly include all studies without further model modifications. However, this approach may be an inadequate treatment of such collections for a number of reasons. For example, the effect sizes of very large observational studies with correspondingly very small standard errors may swamp any evidence provided by the randomized trials (Reeves et al. 2011). This feature would be particularly troubling if the nonrandomized study estimates are biased. Thus, model extensions again permit design-specific differences in treatment effect means.

Prevost et al. (2000) used an approach similar to the grouped random-effects model proposed by Larose & Dey (1997) to posit random differences in design-specific average effects, leading to an extra level in the hierarchy as follows:

$$\hat{\theta}_{jk} \sim \mathrm{N}(\theta_{jk}, \hat{\sigma}_{jk}^2),$$
$$\theta_{jk} \sim \mathrm{N}(\theta_k, \tau_k^2),$$
$$\theta_k \sim \mathrm{N}(\theta, \upsilon^2),$$

where θ_k is the design-specific effect for design type k ($k = 1$ for randomized studies and $k = 2$ for nonrandomized studies). As in traditional random-effects meta-analysis, the goal is to estimate the overall effect θ. A Bayesian approach to parameter estimation is arguably necessitated by the difficulty of estimating the variance υ^2 with only two design types. As Prevost et al. (2000) note, the results are sensitive to the prior distribution for this parameter.

Prevost et al. (2000) demonstrated their method in an analysis of the effect of breast cancer screening promotion on mortality. They estimated a design-averaged relative risk in favor of

screening promotion but noted that (*a*) the effect estimate for the randomized studies was slightly smaller than that for the nonrandomized studies, and (*b*) the variability of this estimate was also slightly smaller than that for the nonrandomized studies. These trends were confirmed by Deeks et al. (2003), who found that treatment effects in observational studies tend to be more variable than those of their randomized counterparts.

Although this analysis was among the earliest examples of combining studies with different designs, few have used it in practice. Grines et al. (2008) used the three-level model directly in their study of the comparative effectiveness of thrombectomy and percutaneous coronary intervention, Sampath et al. (2007) studied the use of loop diuretics for acute renal failure, and McCarron et al. (2010) examined abdominal aortic aneurysms. The latter two studies also implemented model extensions to adjust for perceived data quality.

3.3. Discounting Possibly Low-Quality Data

Rather than embracing the philosophy of including all available relevant studies (e.g., via the three-level hierarchical model), some believe that only the "best" studies (or those of "sufficient" quality) should be included in an analysis (Ades & Sutton 2006, Welton et al. 2009, Higgins et al. 2013). In designed research syntheses, such decisions about inclusion are based on quality scoring systems, such as the Cochrane Collaboration's tool for assessing randomized trials (Higgins et al. 2011) or similar tools for observational studies (Deeks et al. 2003).

A less extreme approach is to include all available studies but to downweight those that are considered to be of lower quality (e.g., nonrandomized designs, which appear lower on the evidence pyramid). Many researchers have suggested methods of implementing such quality weighting. For example, Bérare & Bravo (1998) propose a simple quality score weight to inflate study variance via multiplication. [This idea is very similar to utilization of a power prior distribution for the effect size (Ibrahim & Chen 2000).] Spiegelhalter & Best (2003) flesh out this idea in a model-based context via a very general model that is particularly relevant to combining randomized and nonrandomized data, as it separates external bias (δ_E) from internal bias (δ_I) for each study:

$$\theta_j = \theta + \delta_{Ej} + \delta_{Ij} \tag{1}$$

$$\sim N\left(\theta, \tau_E^2 + \tau_{Ij}^2\right) = N\left(\theta, \frac{\tau_E^2}{q_j}\right), \tag{2}$$

where

$$\delta_{Ej} \sim N\left(0, \tau_E^2\right),$$

$$\delta_{Ij} \sim N\left(0, \tau_{Ij}^2\right),$$

and the quality weight $q_j = (\tau_E^2)/(\tau_E^2 + \tau_{Ij}^2)$ is the fraction of the total variance due to external variability.

As separate information about the random-effects variance τ_E^2 and quality weights $\{q_j\}_{j=1}^J$ in the relevant studies is limited, Spiegelhalter & Best (2003) propose informative prior distributions based on similar meta-analyses or expert opinion. For example, in their assessment of the cost-effectiveness of hip prostheses based on an RCT, a registry, and a case series, Spiegelhalter & Best (2003) used quality weights of 1, 0.5, and 0.2, respectively. These weights reflect an a priori specification that the variance of the case study is inflated fivefold to reflect the true uncertainty in the evidence about the "true" parameter from this "low-quality" design. They also suggest using sensitivity analyses to mitigate the strong influence of these priors.

An alternative that allows nonzero means for the distributions of the bias parameters is to use meta-epidemiological studies to specify prior distributions. Several authors have attempted to use

collections of research syntheses to evaluate the magnitude of biases in randomized trials, for example, as classified by the domains in a quality or risk-of-bias classification scheme (see, e.g., Sterne et al. 2002, Siersma et al. 2007). A more recent proposal by Welton et al. (2009) directly uses these estimates to correct for bias in meta-analysis.

Specifically, Welton et al. (2009) built upon a simplified version of a model proposed by Siersma et al. (2007) to create a meta-regression that includes a single bias indicator:

$$\hat{\theta}_j \sim \mathrm{N}(\theta + \beta_j X_j, \hat{\sigma}^2),$$

where X_j is an indicator of high risk of bias. Thus, the mean for each high-risk-of-bias study effect is randomly offset from the true mean, θ, which is defined as the mean of the low-risk-of-bias study effects. The indicator-based adjustment β_j allows unbiased estimation of this mean. A hierarchy of prior distributions for the random offsets, β_j, reflect between-meta-analysis variability in bias as well as the uncertainty in the overall average bias. This, in turn, effectively downweights the contributions of the studies at high risk of bias. For example, adding ten high-risk trials to an analysis provides a gain in precision that is equivalent to adding between one and six low-risk trials, depending on the hierarchical variability (Welton et al. 2009).

As Welton et al. (2009) note, we could in theory include an observational study in an analysis by defining it as having high risk of bias. They did not attempt such an analysis, however, and suggest that collecting evidence to specify empirically based prior distributions for the bias parameters would be difficult. [Ryan et al. (2012) do use a large collection of observational studies, but they focus on identifying preferable statistical analysis methods rather than reweighting.] Unfortunately, currently available evidence does not support the use of such empirical priors, even for RCT-only analyses. In their study of 148 meta-analyses of randomized trials, Wood et al. (2008) found that the magnitude of biases can vary across different types of studies. For example, they found that biases due to inadequate allocation concealment were larger for subjective outcomes than for objective ones. They also found inconsistencies in the direction of the bias that could have resulted from other study differences.

Welton et al. (2009) demonstrate that the model-based approach can be easily extended to incorporate more sources of variability (e.g., by allowing the θ_j to vary randomly or by allowing the between-study variance to vary across meta-analyses) and more sources of bias [e.g., by assuming that indicators of high risk of bias in multiple domains linearly contribute to the mean; also presented in articles by Siersma et al. (2007) and Turner et al. (2009)]. Turner et al. (2012b) used a version of this quality-based adjustment to analyze data from one randomized and nine nonrandomized studies of a treatment routinely administered to pregnant women to prevent undesirable immune responses during subsequent pregnancies.

Despite the relative ease of specification, models including quality indicators have some major drawbacks. Detailed quality information is often not available, and the intercept in the meta-regression model may not be jointly estimable with the regression coefficients (Higgins et al. 2009). In addition, limited numbers of studies may make estimation for extended models difficult and highly dependent on assumptions. To avoid these disadvantages, it is tempting to collapse multiple dimensions of bias into a single quality score that can be used as a single regression coefficient. Greenland & O'Rourke (2001) provide detailed discussion of this use of quality weights for randomized studies, noting that biases related to quality are not necessarily all in the same direction, much less additive or even linear.

More complex linear bias models are also possible. Dias et al. (2010) propose a mixed treatment comparison meta-analysis similar to that done by Welton et al. (2009). Their approach includes random-effects bias parameters with normal probability distributions that have a

treatment-by-comparison-specific mean. The authors suggested further extensions to include multiple study designs, but they did not explicitly describe or implement such a method.

A few authors have implemented a linear model approach to bias adjustment in analyses of collections including both randomized and nonrandomized studies. In their study of loop diuretics and acute renal failure, Sampath et al. (2007) found moderate evidence that quality score and study completion year were important effect moderators that increase the effect size. No evidence for effect moderation was found for the average age, gender, or control arm risk. McCarron et al. (2010) used imbalance between the treatment arms on three variables as covariates in their study of treatments for abdominal aortic aneurysms. This approach requires that there are a moderate number of imbalanced variables to consider and that arm-specific information is available. If individual-level data were available, bias adjustment via methods such as propensity score adjustments would help to overcome the dimensionality problem. However, analysts seem to more often synthesize randomized and nonrandomized studies separately (see, e.g., Mak et al. 2009, Chowdhury et al. 2014).

4. GENERAL BIAS MODELS

The linear approaches to bias addressed in the previous section can be thought of as special cases of more general bias models. In particular, they are closely tied to the response surface modeling philosophy for meta-analysis that Rubin (1990) proposed as an alternative to traditional meta-analyses focused on simply averaging high-quality study-specific effect sizes (Greenland & O'Rourke 2001). To form the surface, subject-specific effect moderators that affect external validity or generalizability bias can be thought of as being located along one set of axes, and the study-specific effect moderators that affect internal validity, or selection bias (including randomization), can be thought of as being located along a second set of axes. The response surface on the third axis (or set of axes) reflects the expected treatment effect at each point in the two-dimensional subject $\times$ study space. In this framework, the average treatment effect of interest could be identified by a weighted average (or integral) across a subject-specific subspace at a single "ideal" cross-section along the study-specific set of axes. As **Figure 1** shows, the response surface approach is closer to the Spiegelhalter & Best (2003) bias-adjusted model than to the Prevost et al. (2000) three-level average model, where the treatment effect is defined to be the average of the response surface over both the subject and study dimensions.

The simplest response surface models can be created via extensions of the three-level model such as those that Prevost et al. (2000) used to assess the effect of breast cancer screening on mortality. Notably, the authors analyzed data for younger and older women separately. The strikingly different effect estimates suggest that age moderates the effect of cancer screening encouragement. Such moderation can be incorporated into an overarching model for both groups of women and both study design types via an age group indicator covariate while maintaining design-specific variance parameters:

$$\begin{aligned}
\hat{\theta}_{jk} &= \theta + \alpha X_{jk} + \delta_k + \epsilon_{jk}, \\
\delta_k &\overset{\text{ind}}{\sim} \mathrm{N}(0, \upsilon^2), \\
\epsilon_{jk} &\overset{\text{ind}}{\sim} \mathrm{N}(0, \tau_k^2 + \hat{\sigma}_{jk}^2),
\end{aligned}$$

where X_{jk} is an indicator for age group (Prevost et al. 2000). The authors were very thoughtful about specifying design-specific variance components, rather than the tempting but difficult to justify assumption that τ_k^2 is equivalent for all design types k (Turner et al. 2012). The effect of interest was for the older women (θ, i.e., for $X = 0$).

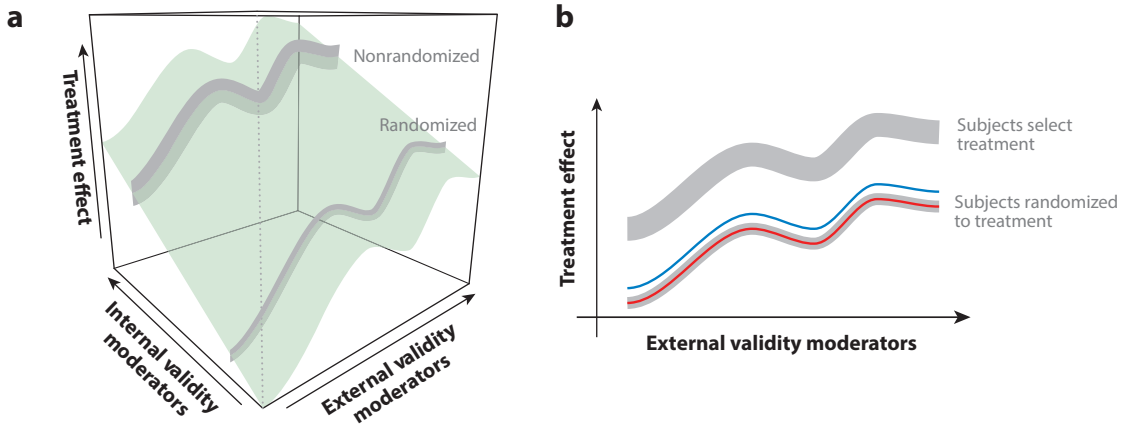

Figure 1

Relationship between response surface, three-level linear models, and bias-adjusted models. (*a*) A hypothetical response surface. Gray lines represent two cross-sections at the internal validity design points, "randomized" and "nonrandomized," where line thickness reflects the precision of each design. (*b*) The same cross-sections as in panel *a* projected in two dimensions. The red line represents the effect surface of interest if the "randomized" design was considered ideal. A single average effect estimate could be a single weighted integral over this line, where the weights are specified by the target population. The blue line represents the effect surface that averages between the two internal validity design points, weighted by the inverse precision. For well-matched studies, the effect estimate from a three-level linear model would be a weighted integral of this surface, where the weights are determined by the study designs. Similarly, the effect estimate from a bias-adjusted model would be a weighted integral of the red line.

This example can be reconceived as a coarse response surface consisting of four points corresponding to the age × design combinations, each of which has corresponding data available for direct estimation. Assumptions about smoothness of the response surface and variance structure would imply a joint model such as that presented by Prevost et al. (2009). The response surface idea could also be used to estimate an average effect across age groups, where the effect size would be a weighted average of the age-specific effect estimates on the "randomized" cross-section of the surface. Of course, if more specific data were available, this type of regression model could also incorporate more study features based on population, intervention, or outcome.

The response surface idea is more powerful in more common situations in which randomized evidence is not available for all of the subject-specific design points. For example, many randomized studies exclude relevant segments of the population for efficiency or ethicality. In this case, the response surface can be used to extrapolate to the design point of interest. This idea is incorporated in cross design synthesis.

4.1. Cross Design Synthesis

An idea that is similar to (but more well developed than) the response surface was promoted by the US General Accounting Office (GAO) (1992) as part of its approach to jointly using experimental and observational evidence in a cross design synthesis (CDS). In its entirety, CDS is a very general four-step methodology:

Step 1 Assess external validity of randomized studies.
Step 2 Assess internal validity of observational studies.
Step 3 Adjust all studies for internal/external validity.
Step 4 Combine adjusted results within and across designs.

Table 1 Synthesis Framework: Primary and Secondary Dimensions of Stratification[a]

	Primary dimension: type of design	
Secondary dimension: coverage of patient groups in randomized studies[b]	**Results of randomized studies: Stratum 1**	**Database analyses: Stratum 2**
Covered in randomized studies (e.g., whites)	Stratum 1a	Stratum 2a
Not covered in randomized studies (e.g., blacks, and other minorities)	Stratum 1b (empty)	Stratum 2b

[a]Reprinted from GAO (1992), table 4.2.

[b]Assumes that existing database analyses cover all patient groups.

The first two steps of a CDS roughly correspond to the two dimensions of the response surface described by Rubin (1990). Step 1 focuses on the subject dimension, and Step 2 focuses on the study dimension. On the basis of these assessments, the GAO approaches Steps 3 and 4 by following Hlatky (1991) in partitioning the design space into four strata based on the study type and randomized study inclusion criteria (see **Table 1**). This stratification is again a coarse representation of a response surface and, again, corresponds to the Prevost et al. (2000) example had no randomized data been available for young women. In this simple CDS, by design, one stratum has no direct evidence, so we must extrapolate to estimate an effect size in that stratum. Following the response surface idea, an estimate of the average treatment effect for the whole population (all ages of women) is a weighted average of the effect size estimates in the "randomized evidence" stratum (**Table 1**, left column). Kaizar (2011) showed that the simple linearly extrapolated estimator is unbiased whenever the bias due to poor internal validity in the observational studies is separate from the inclusion criteria. That is, the estimator is unbiased whenever the bias due to weak external validity in the randomized studies can be estimated without bias, using only data in the observational study strata.

The CDS framework is particularly notable for its explicit treatment of the two dimensions of internal and external validity in a manner that clearly exploits the complementary strengths of randomized designs (which usually have strong internal validity) and observational designs (which usually have strong external validity). To my knowledge, however, no complete practical application of CDS has been published. In fact, the GAO itself used the first three steps of the CDS methodology to examine breast conservation versus mastectomy in the treatment of breast cancer, but it did not complete the fourth and final step: synthesizing the studies (see sidebar, Case Study: US General Accounting Office Cross Design Synthesis Comparing Mastectomy and Breast Conservation Therapy).

4.2. Parametric Bias Models

CDS and response surface methodology can also be viewed as special cases of much broader approaches to bias. Ades & Sutton (2006) provide an in-depth review of general approaches to these types of models under the term "multiparameter evidence synthesis." This subsection reviews some of the major contributors to the literature on parametric bias models, with particular emphasis on applications that relate to the joint analysis of randomized and nonrandomized study data.

One of the earliest general bias models was the confidence profile method (CPM) proposed by Eddy and colleagues (Eddy 1987, 1989; Eddy et al. 1992), which is a methodology for precisely defining questions of interest, relating those questions to available data via probability models, and estimating relevant parameters in those models. In particular, the CPM promotes the use of influence diagrams (graphs) to organize data and clearly identify biases for each study, to create probability models based on the diagrams, and to estimate all of the parameters simultaneously via

CASE STUDY: US GENERAL ACCOUNTING OFFICE CROSS DESIGN SYNTHESIS COMPARING MASTECTOMY AND BREAST CONSERVATION THERAPY

Five relevant RCTs suggest that breast conservation therapy (BCT) is no less effective than mastectomy. There is, however, a potential for considerable treatment heterogeneity that might induce different treatment effects in less-controlled natural practice (e.g., practice in which doctors decide both which patients are good candidates for BCT and how much tissue to remove). Thus, in their 1994 study, the GAO turned to observational data collected in the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) database of cancer patients to examine the generalizability of the RCT results. The GAO focused on creating and analyzing the SEER data to preserve the validity of cross design comparisons.

1. External validity: The GAO included only SEER patients whose characteristics (e.g., tumor size and age) and treatment (e.g., BCT including lumpectomy, nodal dissection, and radiation) were comparable to those included in the RCTs.
2. Internal validity: The GAO compared survival of patients only within propensity score quintiles based on demographic characteristics and tumor size. (Although no quintile showed a statistically significant difference in treatment, there was a trend suggesting that treatment was less effective for those patients who tended to receive BCT less often.)

In its main analysis, which compared treatment effects estimated in the two study designs, the GAO found the SEER-based effect (odds ratio) to be approximately 0.06 smaller than single-site RCTs and 0.11 smaller than multisite RCTs. That is, both designs consistently showed no evidence that BCT is inferior to mastectomy. Because biases due to inadequate adjustment for external and internal validities could potentially offset any true effect differences between the study designs, however, the GAO also conducted a sensitivity analysis that compared the combined treatment survival rates across designs. This analysis showed a 4.3% observed difference in five-year survival rates between the SEER and single-center RCTs, suggesting the possibility of some uncontrolled patient differences between the two study designs.

Using these analyses, the GAO successfully addressed their main questions regarding the generalizability of RCT-based hypothesis tests by comparing estimates between appropriately adjusted analyses in studies with two designs. The GAO did not, however, take the additional step to complete a full CDS by formally jointly modeling all of the data to estimate relevant population quantities.

maximum likelihood or Bayesian methods. [In the event that calculating a simultaneous solution is intractable, Eddy et al. (1992) suggest using a sequential analysis.] In the creation of diagrams and distributions, the CPM makes a clear distinction between the functional parameters for which the data provide direct information (e.g., means that reflect unideal study features), and the true parameter(s) of interest (e.g., the average treatment effect). The functional parameters can often be characterized as a function of the true parameters and nuisance bias parameters.

As an example, consider a study in which 100 λ% of participants are lost to follow-up. If θ_d is the mean outcome for the subjects that drop out, θ_c is the corresponding mean for subjects that complete the study, and θ is the mean outcome across the whole population (the parameter of interest), then we can see that for normally distributed outcomes with constant variance,

$$\theta_c = (\theta - \lambda\theta_d)/(1 - \lambda), \text{ and}$$
$$\overline{Y_c} \sim \text{Normal}(\theta_c, \sigma^2/n_c),$$

where $\overline{Y_c}$ is the observed average outcome for the subjects that completed the study. The goal is to estimate the true parameter θ, but the lack of direct information about the nuisance parameter θ_d precludes standard frequentist likelihood-based estimation. Instead, we must either augment our data with auxiliary sources relevant to θ_d or rely on a prior distribution (based on reasonable assumptions or elicited expert opinions).

Eddy et al. (1992, p. 309) recognized that construction of a model for all of the evidence that incorporates complex biases can be sensitive to many modeling decisions, and they recommend using sensitivity analyses to "explore the impact of uncertainties, assumptions, and judgments." In particular, they suggest using prior distributions for all parameters (i.e., a full Bayesian analysis) and repeating the analysis with different model assumptions. The CPM can address random study-to-study variation via random-effects models for the so-called true parameters, much like the three-level hierarchy described in Section 3.3 does.

Many other authors have used quite similar approaches to modeling biases using several data sources. For example, Greenland (2005, 2009) suggests multiple bias models, focusing on methods for estimation based on sensitivity analyses and discussing technical issues that arise in models with high-dimensional bias parameters. He demonstrates his methodology using case-control studies, for which he indirectly uses auxiliary information to create prior distributions for the bias parameters. Wolpert & Mengersen (2004) promote methods using adjusted likelihoods, demonstrating the use of auxiliary data to provide direct evidence about nuisance bias parameters such as measurement error. Molitor et al. (2009) use Bayesian graphical models to combine several observational studies in order to overcome limitations due to missing data.

In theory, models such as these can correct all types of bias (either empirically by estimating bias parameters or by using a sensitivity analysis), including the different biases typically seen in observational and randomized studies. In practice, however, these bias-correcting methods focus on observational data and corrections for selection biases; they typically do not include both randomized and observational data.

4.3. Generalizability

Generalizability methods are unique in that they nearly uniformly require a synthesis of both observational and randomized data. Their focus on carefully defining the population over which the treatment effect will be estimated, and adjusting the estimates to apply to this population, also sets them apart from most causal modeling. When the number of variables that define the target population is small, standard survey reweighting techniques can be used to reweight the randomized participant data to match the observed variable distributions in the nonrandomized study. For larger numbers of variables, Cole & Stuart (2010) propose using propensity scores to accomplish this reweighting in a single dimension. They demonstrate their method by reweighting participants in a study of antiretroviral therapy among HIV-infected US residents to match the age, sex, and race of the HIV-infected US population, as estimated by the Centers for Disease Control and Prevention (CDC) via observational data.

Unfortunately, this approach is limited by its assumption that all segments of the target population are represented in the study sample. Because randomized trial recruitment criteria are often designed for efficiency and ethicality rather than for generalizability, however, this assumption may not hold. For example, HIV investigators who conducted a trial reanalyzed by Cole & Stuart (2010) only recruited subjects with low CD4 cell counts. If this variable is an effect moderator, any reweighting scheme will result in a partial adjustment (Cole & Stuart 2010). A complete adjustment would be possible only through extrapolation, such as via CDS.

Randomization bias: Changes in the counterfactual outcome of an individual (and thus the estimate of a treatment effect) owing to random assignment to an undesired treatment

When appropriate data for reweighting are not available, it is still prudent to assess the potential extent of study generalizability. Many authors compare baseline characteristics and outcomes between randomized and observational studies, either individually (e.g., Stevens et al. 2007, Greenhouse et al. 2008) or by comparing estimated propensities for enrolling in a trial (e.g., Stuart et al. 2011). If only summaries of the randomized data analysis are available, researchers can at least raise concerns about the potentially limited generalizability of a randomized study owing to inclusion criteria for recruitment by estimating the portion of an observational data set that would have been ineligible to participate (e.g., Fortin et al. 2006, Humphreys et al. 2000, Zimmerman et al. 2004).

Other authors explore limitations to the generalizability of randomized studies that are induced by characteristics that are not directly measurable. For example, many people do not consent to randomization even if they are eligible for and available to enroll in a randomized study. Marcus (1997) proposed a method to examine the effect of these unmeasured selection forces, or nonconsent bias, by comparing randomized trial participants with a nonrandomized registry of patients who were asked to participate in the trial but declined to consent. She used propensity score matching (based on estimated probabilities of belonging to the randomized group) in a study of pharmaceutical versus surgical treatments for otitis media and found modest nonconsent bias indicating that the randomized study effect estimate was too large. In addition, randomization bias can be examined via comprehensive cohorts (in which only participants who are agnostic about treatment choice are randomized) and two-stage randomized designs (in which members of one randomly selected group are allowed to choose their own treatment, whereas those in the other group are randomly assigned to active or control treatment). In a review of studies that used these designs in the medical literature, King et al. (2005) found little evidence for randomization bias large enough to be practically important.

4.4. Chains of Evidence

Another dimension of generalizability concerns the outcome measures of interest. Prospective randomized trials measure many outcomes (such as long-term outcomes or rare events) less efficiently than an observational study would. For example, we would expect a prospective study of 10-year survival to take at least 10 years to complete, whereas we could retrospectively analyze existing data in a matter of weeks. Thus, prospective studies often utilize surrogate outcomes. In such cases, analysts may construct a chain of evidence that leads from the interventions through the surrogate outcome to the outcome of interest. The same general models for bias correction presented above can be repurposed to estimate treatment effects in chain of evidence situations.

For example, Eddy et al. (1992) present a hypothetical use of the CPM for a chain of evidence in which either experimental or observational data are used to estimate parameters for different links in the chain. They also analyze real data related to the effect of tissue-type plasminogen activator (t-PA) on stroke survival, for which most links were estimated using randomized data, but evidence regarding the relationship between intermediate outcome (reperfusion) and outcome of interest (survival) was taken from the control arm of a randomized trial. Ratcliffe et al. (1998) also rely on a chain of evidence modeled via the CPM to estimate the cost-effectiveness of several methods of preventing mother-to-infant transmission of HIV. Most links in this chain were estimated via observational data, but the effect of oral zidovudine in transmission prevention was estimated from a randomized trial.

Different uses of chains of evidence also build on CPM ideas. Epstein et al. (2013) were able to use a randomized trial to learn about the effect of different monitoring tests to prevent diabetic adverse events, but these data contained no direct information about the mechanism of the

treatment. The authors were able to infer properties of the mechanism by incorporating evidence regarding some intermediary outcomes from an observational study. In this spirit, estimating a causal network of variables may assist in better understanding the relationships among a collection of potential moderators and mediators. A substantial literature describes such network discovery based on Bayesian methods and observational data. Cooper & Yoo (1999) and Yoo (2012) demonstrate extensions to these methods that incorporate both randomized and observational data.

Probability of causation: The probability that an intervention is necessary for an outcome to occur

Pearl (2009) also suggests using CPM ideas in a hypothetical estimation of a lower bound for the probability of causation for a treatment. He relies on the strong internal validity of experimental data to estimate a counterfactual mean and on the external validity of observational data to estimate population probabilities. In light of the methods discussed in this section, Pearl's claim that both randomized and nonrandomized designs are necessary for such estimation is quite strong. Nevertheless, this hypothetical example crisply demonstrates the value of joint analysis of data from multiple types of designs.

We must exercise extreme caution in using different sources of evidence to estimate different links in a chain of evidence, however, as these analyses are highly dependent on model assumptions. Similar phenomena have long been recognized in available case approaches to analyses with missing data. For example, we may estimate a variance-covariance matrix with a matrix constructed of pairwise sample covariances, calculating each element using all individuals for whom that pair of variables was measured. Because the elements of the matrix are based on different subsamples of the data, the resulting estimated covariance matrix may not be positive definite (Little & Rubin 2002). In fact, because the mechanisms that caused the missing data imply that each covariance parameter may be conditional on different population properties, a theoretical variance-covariance matrix constructed in this manner is also not guaranteed to be positive definite or to represent the variance-covariance matrix for the intended full sample. The chain-of-evidence analog is that the mechanisms or study designs that created the different data sets (i.e., the known or latent participant, intervention, environment, and outcome characteristics that differ across studies) may lead to discordant or incongruent parameter estimates. As Pearl (2009, p. 303) notes, the combination of estimates he suggests is valid only when the participants in both studies "were sampled properly from the population at large." That is, the estimate is valid only if the assumed models for external validity are correct.

5. DISCUSSION

In the current climate of increasing demand for timely inference about treatment effects, turning to collections of relevant data with diverse designs may be both more efficient and more robust than relying on the so-called best evidence from a single design. Although some comparisons indicate that randomized and nonrandomized studies provide divergent evidence, careful analyses that precisely consider the definition of the target average treatment effect show markedly more agreement across designs.

Development of statistical methodology to combine information across designs has generally fallen into two categories: (*a*) those that use variables that are potentially related to bias (design elements or quality scores) to adjust the mean of a model for study-specific effect sizes (Section 3), and (*b*) those that consider more general models for bias-generating mechanisms and chains or networks of evidence that synthesize evidence on different parts of a complete causal picture (Section 4). Neither category has benefitted from an overabundance of methods development or practical application specifically related to incorporating studies with both randomized and nonrandomized designs. The linear approaches have received more attention, likely because of

their simpler implementation with less need for application-specific customization and detailed individual-level data.

Even if data were more available, however, general bias methods are not developed to the point at which rich data and plentiful manpower could be easily exploited. Within a few sound general frameworks, several authors have described specific methods for overcoming some specific insults to external or internal validity. Unfortunately, many more threats to sound causal conclusions exist, and specific methods for overcoming them remain undeveloped. One particularly inviting area is methods to jointly address internal and external biases. As reviewed in this article, these two dimensions of validity have largely been addressed separately. Even though in theory these dimensions could be jointly applied, the details of doing so must be worked out and the feasibility demonstrated with examples. Looking forward, considerations regarding research synthesis should also be incorporated into the design of both randomized and nonrandomized studies. The current literature is lacking in such guidance.

Joint analysis across designs would also benefit from further development of several areas of statistics not reviewed here but necessary to support practical joint data analysis. For example, methods to harmonize variables measured differently in different studies are essential for the practical joint analysis of many studies of the same type, but such methods are arguably even more important for synthesizing collections of studies designed within different traditions. Methods for exploring treatment effect heterogeneity are another example. The random-effects models discussed in Sections 3 and 4 are usually justified via the recognition of study-to-study variability in true study-specific treatment effect. However, the development of methods to exploit this variability in order to estimate more personalized effect sizes has only recently begun in earnest.

Much groundwork has already been laid for sound causal inference based on collections of studies that incorporate both randomized and nonrandomized study designs, but the amount of work still needed to complete the development of detailed methods suitable for practical application is considerable. As the availability of both randomized and nonrandomized data continues to grow, the joint analysis of all available relevant data will play a key role in fulfilling the increasing demand for treatment effect estimates for use in evidence-based decision making.

SUMMARY POINTS

1. The exact definition of the treatment effect of interest determines the relative value of various study designs and statistical methods. Researchers wishing to make causal conclusions must be precise in defining the exact treatment effect of interest and in specifying any assumptions about treatment effect heterogeneity.
2. Estimates and tests of average treatment effects vary from study to study and may appear to correlate with the type of study design. But, variation in the precise definition of and assumptions about average treatment effect across the designs is likely more responsible for such observed correlation than are any inherent strengths or weaknesses of the study designs.
3. Researchers have historically placed greater value on the strong internal validity typical of randomized designs than on the strong external validity typical of observational designs. The view that the two designs have complementary strengths will yield better statistical inference in many cases.

4. Joint analyses of data from randomized studies and from observational studies can provide more information about a treatment effect than two separate analyses would. In such cases, each study may relate directly to the effect (with some possible bias) or may contribute information about some pieces of a chain of evidence leading from treatments to outcomes.
5. Researchers have created a variety of methods to jointly analyze data from collections of studies with multiple design types. Some methods are based on weighted averages of the study-specific effect estimates; some more general approaches are designed to estimate the results of an ideal internally and externally valid study via extrapolation, interpolation, and averaging.
6. Developing and improving methods to jointly model data from collections of differently designed studies, as well as methods to address specific sources of selection and generalizability bias, may in many cases reduce the cost and increase the timeliness of robust inference about treatment effects.

FUTURE ISSUES

1. Further development of prospective designs for multipart studies that include both randomization and natural observation will facilitate more efficient, accurate, and generalizable inference.
2. More nuanced use of existing data to plan future studies will facilitate more efficient, unbiased, and generalizable inference.
3. Whereas separate ideas for mitigating selection bias and generalizability bias have been utilized together in single analyses, single methods that jointly target both types of bias may improve estimation.
4. Improving methods for variable harmonization in complex, large-scale data sets will make joint analyses more practical.
5. The goal of causal analyses will continue to move from the estimation of average treatment effects in broad populations to the estimation of treatment effects that are specific to subgroups or individuals. Robust statistical strategies must be developed to support this shift.
6. Publishing practical recommendations and more examples would promote precise and accurate estimation of relevant treatment effects via the joint analysis of diverse data sets.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank an anonymous referee for providing extremely helpful suggestions.

LITERATURE CITED

Ades AE, Sutton AJ. 2006. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *J. R. Stat. Soc. A* 169(1):5–35

Amatya A, Bhaumik DK, Normand S-L, Greenhouse J, Kaizar E, et al. 2015. Likelihood-based random effect meta-analysis of binary events. *J. Biopharm. Stat.* In press. doi: 10.1080/10543406.2014.920348

Begg CB, Pilote L. 1991. A model for incorporating historical controls into a meta-analysis. *Biometrics* 47:899–906

Benson K, Hartz AJ. 2000. A comparison of observational studies and randomized controlled trials. *N. Engl. J. Med.* 342:1878–86

Bérare A, Bravo G. 1998. Combining studies using effect sizes and quality scores: application to bone loss in postmenopausal women. *J. Clin. Epidemiol.* 51:801–7

Bhaumik DK, Amatya A, Normand S-L, Greenhouse J, Kaizar E, et al. 2012. Meta-analysis of rare binary adverse event data. *J. Am. Stat. Assoc.* 107(498):555–67

Black N. 1996. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 312(7040):1215–18

Bridge JA, Axelson DA. 2008. The contribution of pharmacoepidemiology to the antidepressant-suicidality debate in children and adolescents. *Int. Rev. Psychiatry* 20(2):209–14

Chowdhury R, Kunutsor S, Vitezova A, Oliver-Williams C, Chowdhury S, et al. 2014. Vitamin D and risk of cause specific death: systematic review and meta-analysis of observational cohort and randomised intervention studies. *BMJ* 348:g1903

Cohen AM, Stavri PZ, Hersh WR. 2004. A categorization and analysis of the criticisms of evidence-based medicine. *Int. J. Med. Inform.* 73:35–43

Cole SR, Stuart EA. 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG-320 trial. *Am. J. Epidemiol.* 172:107–15

Concato J, Shah N, Horwitz RI. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* 342(25):1887–92

Cook TD, Shadish WR, Wong VC. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J. Policy Anal. Manag.* 27(4):724–50

Cooper GF, Yoo C. 1999. Causal discovery from a mixture of experimental and observational data. *Proc. 15th Conf. Uncertain. Artif. Intell.*, Stockholm, Swed., pp. 116–25. San Francisco: Morgan Kaufmann

Cooper WO, Callahan ST, Shintani A, Fuchs DC, Shelton RC, et al. 2014. Antidepressants and suicide attempts in children. *Pediatrics* 133(2):204–10

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, et al. 2003. Evaluating non-randomised intervention studies. *Health Technol. Assess.* 7(27):1–173

Dias S, Sutton AJ, Welton NJ, Ades AE. 2013. Evidence synthesis for decision making 3: heterogeneity—subgroups, meta-regression, bias, and bias-adjustment. *Med. Decis. Making* 33:618–40

Dias S, Welton NJ, Marinho VCC, Salanti G, Higgins JPT, Ades AE. 2010. Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. *J. R. Stat. Soc. A* 173(3):613–29

Eddy DM. 1987. The use of confidence profiles to assess tissue-type plasminogen activator. In *Acute Coronary Care 1987*, ed. RM Califf, GS Wagner, pp. 89–110. Boston: Nijhoff

Eddy DM. 1989. The confidence profile method: a Bayesian method for assessing health technologies. *Oper. Res.* 37:210–28

Eddy DM, Hasselblad V, Shachter RD. 1992. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. Boston: Academic

Epstein D, Mochón LG, Espín J, Soares MO. 2013. Use of multiparameter evidence synthesis to assess the appropriateness of data and structure in decision models. *Med. Decis. Making* 33:715–30

Fortin M, Dionne J, Phiho G, Gignac J, Almirall J, Lapointe L. 2006. Randomized controlled trials: Do they have external validity for patients with multiple comorbidities? *Ann. Fam. Med.* 4:104–8

Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29

GAO (US Gen. Account. Office). 1992. *Cross design synthesis: a new strategy for medical effectiveness research*. GAO PEMD-92-18, Washington, DC. **http://www.gao.gov/assets/160/151472.pdf**

GAO (US Gen. Account. Office). 1994. *Breast conservation versus mastectomy: patient survival in day-to-day medical practice and in randomized studies*. GAO PEMD-95-9, Washington, DC. **http://www.gpo.gov/fdsys/pkg/GAOREPORTS-PEMD-95-9/pdf/GAOREPORTS-PEMD-95-9.pdf**

Gibbons RD, Brown CH, Hur K, Marcus SM, Bhaumik DK, et al. 2007. Early evidence on the effects of regulators' suicidality warnings on SSRI prescriptions and suicide in children and adolescents. *Am. J. Psychiatry* 164:1356–63

Gibbons RD, Hur K, Bhaumik DK, Mann JJ. 2006. The relationship between antidepressant prescription rates and rate of early adolescent suicide. *Am. J. Psychiatry* 163:1893–904

Green DP, John P. 2010. Field experiments in comparative politics and policy. *Ann. Am. Acad. Polit. Soc. Sci.* 628:6–10

Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. 2008. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Stat. Med.* 27(11):1801–13

Greenland S. 2005. Multiple-bias modelling for analysis of observational data. *J. R. Stat. Soc. A* 168:267–306

Greenland S. 2009. Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Stat. Sci.* 24(2):195–210

Greenland S, O'Rourke K. 2001. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2:463–71

Grines CL, Nelson TR, Safian RD, Hanzel G, Goldstein JA, Dixon S. 2008. A Bayesian meta-analysis comparing AngioJet® thrombectomy to percutaneous coronary intervention alone in acute myocardial infarction. *J. Interv. Cardiol.* 21:459–82

Grootendorst DC, Jager KJ, Zoccali C, Dekker FW. 2010. Observational studies are complementary to randomized controlled trials. *Nephron Clin. Pract.* 114:c173–77

Hamad TA, Laughren T, Racoosin J. 2006. Suicidality in pediatric patients treated with antidepressant drugs. *Arch. Gen. Psychiatry* 63:332–39

Heckman JJ, Smith JA. 1995. Assessing the case for social experiments. *J. Econ. Perspect.* 9(2):85–110

Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, et al. 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 19:766–79; discussion pp. 780–93

Higgins JPT, Altman DG, Sterne JAC, eds. 2011. Chapter 8: assessing risk of bias in included studies. In *Cochrane Handbook for Systematic Reviews of Interventions*, version 5.1.0 (*updated March 2011*), ed. JPT Higgins, S Green. Cochrane Collab. **http://handbook.cochrane.org**

Higgins JPT, Ramsay C, Reeves BC, Deeks JJ, Shea B, et al. 2013. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Res. Synth. Methods* 4:12–25

Higgins JPT, Thompson SG, Spiegelhalter DJ. 2009. A re-evaluation of random-effects meta-analysis. *J. R. Stat. Soc. A* 172(1):137–59

Hlatky MA. 1991. Using databases to evaluate therapy. *Stat. Med.* 10:647–52

Hlatky MA, Califf RM, Harrell FE, Lee KL, Mark DB, Pryor DB. 1998. Comparison of predictions based on observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery. *J. Am. Coll. Cardiol.* 11(2):237–45

Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, et al. (Oxford Cent. Evid. Based Med. Levels Evid. Work. Group). 2011. *The Oxford 2011 Levels of Evidence*. Oxford Cent. Evid. Based Med. **http://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf**

Humphreys K, Weisner C. 2000. Use of exclusion criteria in selecting research subjects and its effect on the generalizability of alcohol treatment outcome studies. *Am. J. Psychiatry* 157:588–94

Ibrahim JG, Chen M-H. 2000. Power prior distributions for regression models. *Stat. Sci.* 15(1):46–60

Imai K, King G, Stuart EA. 2008. Misunderstandings among experimentalists and observationalists about causal inference. *J. R. Stat. Soc. A* 171(2):481–502

Imai K, van Dyk DA. 2004. Causal inference with general treatment regimes: generalizing the propensity score. *J. Am. Stat. Assoc.* 99(467):854–66

Ioannidis JPA, Chang CQ, Lam TK, Schully SD, Khoury MJ. 2013. The geometric increase in meta-analyses from China in the genomic era. *PLOS ONE* 8(6):e65602

Ioannidis JPA, Haidich A-B, Pappa M, Pantzakis N, Kokori SI, et al. 2001. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 286:821–30

IOM (Institute of Medicine). 2013. *Observational Studies in a Learning Health System: Workshop Summary*. Washington, DC: Nat. Acad. Press

Kaizar EE, Greenhouse JB, Seltman H, Kelleher K. 2006. Do antidepressants cause suicidality in children? A Bayesian meta-analysis. *Clin. Trials* 3(2):73–98

Kaizar EE. 2011. Estimating treatment effect via simple cross design synthesis. *Stat. Med.* 30(25):2986–3009

King M, Nazareth I, Lampe F, Bower P, Chandler M, et al. 2005. Impact of participant and physician intervention preferences on randomized trials: a systematic review. *JAMA* 293(9):1089–99

Larose DT, Dey DK. 1997. Grouped random effects models for Bayesian meta-analysis. *Stat. Med.* 16(16):1817–29

Li Z, Begg CB. 1994. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *J. Am. Stat. Assoc.* 89:1523–27

Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data.* Hoboken, NJ: Wiley. 2nd ed.

Lu CY, Zhang F, Lakoma MD, Madden JM, Rusinak D, et al. 2014. Changes in antidepressant use by young people and suicidal behavior after FDA warnings and media coverage: quasi-experimental study. *BMJ* 348:g3596

MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. 2000. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol. Assess.* 4(34):1–154

Mak A, Cheung MWL, Ho RC-M, Cheak AA-C, Lau CS. 2009. Bisphosphonates and atrial fibrillation: Bayesian meta-analyses of randomized controlled trials and observational studies. *BMC Musculoskelet. Disord.* 10:113

Marcus SM. 1997. Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *J. Clin. Epidemiol.* 50(7):823–28

Marcus SM, Stuart EA, Wang P, Shadish WR, Steiner PM. 2012. Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychol. Methods* 17(2):244–54

McCarron CE, Pullenayegum EM, Thabane L, Goeree R, Tarride J-E. 2010. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Med. Res. Methodol.* 10:64

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control. Clin. Trials* 16:62–73

Molitor N-T, Best N, Jackson C, Richardson S. 2009. Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth-weight and water disinfection by-products. *J. R. Stat. Soc. A* 172(3):615–37

Peto R. 1987. Why do we need systematic overviews of randomized trials? *Stat. Med.* 6:233–40

Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. New York: Cambridge Univ. Press. 2nd ed.

Peinemann F, Tushabe DA, Kleijnen J. 2013. Using multiple types of studies in systematic reviews of health care interventions—a systematic review. *PLOS ONE* 8(12):e85035

Prevost TC, Abrams KR, Jones DR. 2000. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Stat. Med.* 19:3359–76

Ratcliffe J, Ades AE, Gibb D, Sculpher MJ, Briggs AH. 1998. Prevention of mother-to-child transmission of HIV-1 infection: alternative strategies and their cost-effectiveness. *AIDS* 12:1381–88

Reeves BC, Deeks JJ, Higgins JPT, Wells GA. 2011. Chapter 13: including non-randomized studies. In *Cochrane Handbook for Systematic Reviews of Interventions*, version 5.1.0 (*updated March 2011*), ed. JPT Higgins, S Green. Cochrane Collab. **http://handbook.cochrane.org**

Reeves BC, Higgins JPT, Ramsay C, Shea B, Tugwell P, Wells GA. 2013. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Res. Synth. Methods* 4:1–11

Rothwell PM. 2005. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet* 365(9453):82–93

Rubin D. 1990. A new perspective on meta-analysis. In *The Future of Meta-Analysis*, ed. KM Wachter, ML Straff, pp. 155–165. New York: Russell Sage Found.

Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. 2012. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat. Med.* 31:4401–15

Sampath S, Moran JL, Graham PL, Rockliff S, Bersten AD, Abrams KR. 2007. The efficacy of loop diuretics in acute renal failure: assessment using Bayesian evidence synthesis techniques. *Crit. Care Med.* 35(11):2516–24

Shadish WR, Clark MH, Steiner PM. 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J. Am. Stat. Assoc.* 103:1334–44

Shrier I, Boivin J-F, Steele RJ, Platt RW, Furlan A, et al. 2007. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *Am. J. Epidemiol.* 166(10):1203–9

Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. 2007. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. *Stat. Med.* 26:2745–58

Spiegelhalter DJ, Best NG. 2003. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat. Med.* 22:3687–709

Sterne JAC, Jüni P, Schultz KF, Altman DG, Bartlett C, Egger M. 2002. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat. Med.* 21:1513–24

Stevens J, Kelleher K, Greenhouse J, Chen G, Xiang H, et al. 2007. Empirical evaluation of the generalizability of the sample from the multimodal treatment study for ADHD. *Admin. Policy Ment. Health Ment. Health Serv. Res.* 34(3):221–32

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Stat. Soc. A* 174(2):369–86

Teicher MH, Glod C, Cole JO. 1990. Emergence of intense suicidal preoccupation during fluoxetine treatment. *Am. J. Psychiatry* 147:207–10

Thompson SG. 1994. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 309:1351–55

Thompson SG, Higgins JP. 2002. How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* 21(11):1559–73

Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. 2012. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the *Cochrane Database of Systematic Reviews*. *Int. J. Epidemiol.* 41(3):818–27

Turner RM, Lloyd-Jones M, Anumba DOC, Smith GCS, Spiegelhalter DJ, et al. 2012. Routine antenatal anti-D prophylaxis in women who are Rh(D) negative: meta-analyses adjusted for differences in study design and quality. *PLOS ONE* 7(2):e30711

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. 2009. Bias modelling in evidence synthesis. *J. R. Stat. Soc. A* 172:21–47

Valuck RJ, Libby AM, Sills MR, Giese AA, Allen RR. 2004. Antidepressant treatment and risk of suicide attempt by adolescents with major depressive disorder: a propensity-adjusted retrospective cohort study. *CNS Drugs* 18(15):1119–32

Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. 2009. Models for potentially biased evidence in meta-analysis using empirically based priors. *J. R. Stat. Soc. A* 172(1):119–36

Wilks DC, Mander AP, Jebb SA, Thompson SG, Sharp SJ, et al. 2011a. Dietary energy density and adiposity: employing bias adjustments in a meta-analysis of prospective studies. *BMC Public Health* 11:48

Wilks DC, Sharp SJ, Ekelund U, Thompson SG, Mander AP, et al. 2011b. Objectively measured physical activity and fat mass in children: a bias-adjusted meta-analysis of prospective studies. *PLOS ONE* 6(2):e17205

Williams DDR, Garner J. 2002. The case against 'the evidence': a different perspective on evidence-based medicine. *Br. J. Psychiatry* 180:8–12

Wolpert R, Mengersen K. 2004. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Stat. Sci.* 19:450–71

Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, et al. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 336:601

Yoo C. 2012. The Bayesian method for causal discovery of latent-variable models from a mixture of experimental and observational data. *Comput. Stat. Data Anal.* 56:2183–205

Zimmerman M, Chelminski I, Posternak MA. 2004. Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *J. Nerv. Ment. Dis.* 192:87–94