Synthesizing evidence through bias estimation and adjustment

Jasper Cooper*

October 23, 2017

Abstract

I show that it is possible to recover the informative content of evidence that is at risk of bias, even if the location and scale of the bias are completely unknown. The method exploits an assumption that the bias in evidence on an estimand of interest is similar to bias in another source of evidence about a different estimand. The use of empirical estimation procedures to adjust for bias in evidence synthesis presents an advantage over existing approaches, which derive information on bias from experts who may themselves form opinions with unknown systematic error.

*jjc2247@columbia.edu Columbia University.

Contents

Introduction

In principle, the synthesis of all available evidence on an estimand provides high quality inferences because it draws on the entire body of accumulated scientific knowledge. In practice, however, much of the available evidence on an estimand is deemed at risk of unknown and heterogeneous forms of bias. In some cases, the only available evidence on some estimand may be at risk of completely unknown bias. This problem is routinely ignored by researchers.

The ideal synthesis strategy is one in which the informative content of evidence at risk of bias is recovered and used to improve inferences. However, evidence at risk of bias is only informative about an estimand if the information used to learn about bias is not entirely derived from knowledge about the estimand (Gerber, Green, and Kaplan, 2004; Welton et al., 2009). Predominant approaches to bias adjustment in meta analysis attempt to overcome this problem by gathering informative priors on the scale and location of bias terms through expert elicitation (Turner et al., 2009). However, this technique only works if the process through which experts form opinions is completely known. If the opinion formation process is unknown, rather than solving the problem of unknown bias, expert-informed priors simply displace it from the evidence on the estimand to the priors provided by experts.

I propose in this paper an alternative method for synthesizing evidence by empirically estimating and adjusting for bias. A key advantage of this method is that the informative content of evidence at risk of bias is recovered even if the meta-analyst has no prior knowledge about the scale and location of the bias. The method, Bias Estimation Adjusted Meta Analysis (BEAMA), rests on a different core assumption from that of informative unbiased priors: namely, that two or more pieces of evidence on *different* estimands are generated with *correlated* bias. I argue that this assumption is more easily defended than that of the unbiasedness of experts, because it relies on the same grounds upon which the meta-analyst premises her assumption that different sources of evidence encode information on the same estimand, despite evidence being generated on different samples, in different places, at different times. Specifically, the assumption of correlated bias is grounded in the empirical observation of isomorphism in evidence-generating processes.

My key finding is that, if biased data are the only available evidence on an estimand, that evidence can still be perfectly informative about the estimand if the bias is contained in evidence on another, possibly unrelated, estimand. In simulation studies, I show that in certain such instances, biased data on an ancillary estimand is in fact marginally more informative about the primary estimand than *un*biased data on the ancillary estimand. The possibility results are suggestive of a new method for data collection and analysis in meta-analysis of existing studies, and are especially relevant to investigations in which it

is not feasible to gather evidence free from any kind of bias, which arguably characterizes a wide range of research questions. The possibility findings also suggest new ways to collect and analyze possibly biased yet informative expert data.

Section 1 provides a critical appraisal of predominant approaches to bias adjusted meta-analysis. Section 2 formalizes the core possibility results that justify the use of BEAMA as a strategy for data collection and analysis, using simple algebraic examples. Section 3 presents a framework for decomposing additive error that helps provide empirical grounds for the correlated bias assumption, and informs the statistical model developed in section 4. Section 5 investigates the inferential potential and risks in BEAMA through four simulation studies, and section 6 discusses and concludes.

1 Literature Review

Increasingly researchers are in search of appropriate methods to form inferences through the integration of very heterogeneous sources of knowledge (Humphreys and Jacobs, 2015). In the meta-analysis literature, the concept of synthesizing multiple sources of evidence while adjusting for the possible risks of bias in those sources of evidence was first introduced by Eddy, Hasselblad, and Shachter (1990) and Eddy et al. (1992). In their outline for the Confidence Profile Method (CPM), the authors advocated including explicit parameterizations of the evidence-generating process, including nuisance parameters such as bias. According to the logic of CPM, by accounting and adjusting for bias in heterogeneous sources of evidence, meta-analysts could bring to bear the entire accumulated body of evidence onto inferences about an estimand. Around this time, the U.S. Government Accountability Office proposed an even more ambitious use of bias adjustment to make joint inferences incorporating observational and experimental evidence (GAO, 1992). In their proposed method, researchers would leverage the perceived external validity of observational studies and the internal validity of experimental studies by first assessing the external and internal validity of randomized and observational studies, respectively, adjusting for those assessments, and then combining adjusted results within and across designs. However, as Kaizar (2015, 61) points out, "no complete practical application of CDS has been published." The key obstacle in implementing these proposals resides in the need to simultaneously learn about biases and estimands.

At least two studies have characterized this problem formally (Gerber, Green, and Kaplan, 2004; Welton et al., 2009). Both feature a Bayesian meta-analyst who aims to learn about a primary estimand, possesses biased and unbiased information on that estimand, and must choose which evidence to consider. The key insight of these models is that a source of evidence at risk of bias whose location and scale is

entirely unknown provides no information about the estimand of interest. Gerber, Green, and Kaplan (2004) refer to this as the 'illusion of learning' theorem: while intuitively it might appear possible to measure the bias in at-risk pieces of evidence on an estimand through comparison with evidence that is not at-risk of bias on that same estimand, and then to use this measurement to recover the informative content of the at-risk evidence, in practice this is not the case. As long as the researcher knows nothing about the nature of the unobserved bias, then the only informative evidence is the unbiased evidence. The implication of this finding is that even pieces of biased evidence with extremely low sample variance will not contribute new information to a meta-analysis whose unbiased evidence has high sample variance, because the identification of the unknown bias derives entirely from the noisy unbiased evidence.

In order to recover the informative content of evidence at risk of unknown bias, independent information on that bias must be brought to bear on the analysis. Wolpert and Mengersen (2004) presented an approach to this problem that builds on the "quality weighting" approach of Spiegelhalter and Best (2003). In it, the authors specify informative priors about the bias parameters based on their knowledge of the evidence-generating process. In a well-cited example of this informative-prior approach to bias adjusted evidence synthesis, Thompson et al. (2011) conducted a large-scale meta-analysis of the relationship between physical activity and adiposity in children, in which they enumerated six sources of internal bias and four sources of external bias, and elicited the opinions of six statisticians about the scale and location of these different sources of bias in the different studies. This evidence on bias is then incorporated into the inferences about the causal estimand as informative prior information in a Bayesian model with explicit bias parameters. Similar work has been conducted by Darvishian et al. (2014) in their bias adjusted meta analysis of influenza treatments.

Developing informative priors about the unknown bias parameters renders evidence at risk of bias informative. Arguably, however, this approach features a fundamental flaw: it is only valid if expert opinions about bias are themselves unbiased. In other words, the assumption of unbiasedness that was deemed untenable in the meta-analysis and thus prompted the use of a bias adjustment procedure is simply transferred onto the expert data. Many studies have illustrated that expert appraisals are at risk of various forms of cognitive bias (Tversky and Kahneman, 1975; Kynn, 2008). It is therefore risky to assume that expert bias is known to be zero. While all assumptions contain some degree of risk, the assumption that experts are unbiased is very difficult to interrogate in a single meta-analysis using the methods described.

One might assume that the experts' appraisal of bias is itself unbiased because they derived it by comparing unbiased and biased studies on the estimand of interest. Or indeed, one might compare

unbiased evidence about the estimand of interest to the expert's opinions on bias in order to assess the validity of their opinions. Note, however, that both of these cases are examples of the illusion of learning theorem: all information about the unknown bias comes from the unbiased information on the estimand, essentially rendering the experts' opinions uninformative. In sum, approaches that incorporate elicited priors about bias parameters share the same risks of the undadjusted meta-analysis: expert opinions are themselves data at risk of unknown bias, and are only informative insofar as the meta-analyst can make assumptions about the scale and location of that bias.

In this paper, I suggest a different approach that does not require informative priors about the scale and location of bias in order to recover the informative content of evidence at risk of bias. I refer to the approach as Bias Estimation Adjusted Meta-Analysis (BEAMA), because information on bias terms derives from an empirical estimation procedure. In BEAMA, even evidence at risk of bias about whose location and scale the meta-analyst is entirely ignorant can be informative.

The BEAMA approach relies on an alternative assumption that I argue is more easily defended on empirical grounds than the assumption of unbiased priors. Specifically, it employs joint priors on two pieces of evidence on *different* estimands that exploit the belief that the bias works in similar ways across the two sources of evidence. Without knowing the location and scale of the bias in two pieces of evidence, a researcher may have good grounds to assume that the bias is the same or at least similar in both, and expresses this assumption through joint priors that induce correlation in the beliefs about bias. Since the joint priors induce correlation in priors on the bias, I refer to this as the correlated bias assumption.

The grounds for assuming that bias is similar across two pieces of evidence on different estimands are the same that a meta-analyst uses when she assumes that two pieces of evidence pertain to a common estimand, in spite of observable differences in the evidence-generating processes. Specifically, this assumption relies on the observation of isomorphism across evidence-generating procedures. For example, if a study estimates two estimands – say, an estimate of the effect of treatment A vs. treatment B compared to some placebo – it generates two pieces of evidence. Because it uses the same sampling, assignment and measurement procedures to generate those two pieces of evidence, one might reasonably defend the assumption that these two estimates exhibit similar bias, and exploit this assumption through joint priors on the unknown location and scale of the bias in the two studies.

My approach builds off a recent literature on mixed treatment comparisons in networked meta-analysis (Schmitz, Adams, and Walsh, 2013; Efthimiou et al., 2016). An embryonic form of BEAMA was implemented by Dias et al. (2010), who sought to estimate and adjust for bias from non-blinding in clinical trials on the effect of fluoride treatments on caries among children. The approach is also similar to that

used in the network meta-analysis by Chaimani et al. (2013), who conduct a meta-epidemiological study of 32 groups of treatment comparisons comprising 613 trials from different medical fields. Both of these studies have the potential to leverage the correlated bias assumption but do not do so formally. Neither study investigates the conditions that must be met or the potential for inferential improvement that derives from using the correlated bias assumption in a meta-analysis network comparing different estimands. McCarron et al. (2011) do provide a simulation study of a bias estimation and adjustment procedure, but do not investigate the specific technique used in BEAMA. Rather, they assume that bias can be measured and fully accounted for using a single observable measure (the imbalance in age across treatment conditions), which seems unrealistic in practice.[1] In contrast, I provide formal possibility results that show the BEAMA strategy can indeed be used to recover the informative content of biased data, investigate how the assumptions it relies upon can be defended, and provide a series of simulation studies that explicitly investigate its inferential properties within a potential outcomes framework.

2 Bias Estimation Adjusted Meta-Analysis

A researcher wants to learn about the true value of a causal estimand, τ. Suppose that the available evidence on τ is at risk of some unknown bias, β_τ. Bias is here defined as the expected difference between the true value of an unknown parameter and the expected estimate of that unknown parameter (for example, the expected posterior mean).

There exists evidence on an ancillary estimand, ψ, that is also at risk of bias β_ψ. For ease of exposition, I first assume that $\beta_\tau = \beta_\psi = \beta$, such that the bias is identical in evidence on the two estimands (although its true location and scale are unknown). Subsequently in subsection 3, I consider the case in which the biases are not identical but operate similarly to show how a researcher can exploit this belief through priors on the joint distribution of β_τ and β_ψ.

Definition 1. *An **evidence synthesis strategy** is defined by the choice of* $\mathbf{y} \in \mathcal{P}(\mathcal{Y})$*, where* $\mathbf{y}$ *is a vector of evidence,* $\mathcal{Y}$ *is the set of all available evidence, and* $\mathcal{P}(\mathcal{Y})$ *its power set.*

The researcher must choose an evidence synthesis strategy. That is, she must decide to condition her inferences about τ on some combination of the available evidence.

[1] For one thing, it is unlikely that the extent of bias can be measured through observable covariates. For another thing, the problem of observable imbalance in observational studies could presumably be overcome at the study level through the use of covariate adjustment.

2.1 Independence of Estimands

Suppose that the researcher has access to three potential sources of evidence, $\mathcal{Y} = \{y_{\tau+\beta}, y_{\psi+\beta}, y_{\psi}\}$. The first of these, $y_{\tau+\beta}$, is the only direct source of evidence on τ, and is at risk of unknown bias β. The second, $y_{\psi+\beta}$ is on an ancillary estimand that is not of direct interest to the researcher, ψ, and is also deemed at risk of the same kind of unknown bias, β. Finally, there exists unbiased data on the ancillary estimand, y_{ψ}.

To simplify the exposition, assume for now that estimands and bias can only take binary values and that the data-generating process has an additive functional form, such that: $\tau \in \{0, 1\}$, $\psi \in \{0, 1\}$ and $\beta \in \{0, 1\}$; and $y_{\tau+\beta} = \tau + \beta$, $y_{\psi+\beta} = \psi + \beta$, and $y_{\psi} = \psi$.

The researcher must choose an evidence synthesis strategy conditional on her prior beliefs about how likely various potential outcomes are. Let $\Pr(\theta)$ describe one possible vector of beliefs about the probabilities of different events occurring, and Θ the matrix of all possible priors. For example, the researcher might want to use informative priors that put more probability mass on the states of the world in which the unknown bias is 1 than on those in which the bias is 0. In the following, however, assume that the researcher uses 'non-informative' priors, in the sense that they do not contain information about the true scale or location of the bias. This reflects the fact that the researcher is ignorant about the true values of the parameters.

Specifically, the researcher considers only those $\Pr(\theta) \in \Theta$ in which the marginal probability of the different values of the unknown parameters is equal. In other words, she always attributes a prior mean of .5 to each element of the vector of unknown parameters, $\theta = [\tau \quad \psi \quad \beta]$. Although the different values are considered marginally equally likely, the researcher does not always attribute equal probabilities to the *joint* values of the unknown parameters, however. This is the correlated bias assumption.

For example, she might believe that events in which the two estimands have the same value are more likely than ones in which they have different values, because the evidence was generated in a similar way. In that case, although the researcher's marginal priors on the values of the unknown parameters would be equal, her beliefs on the joint probabilities of different parameter values induce correlation in the bias priors. Again, we begin with the extreme case in which the bias is literally identical across studies.

Given her lack of knowledge about the true value of the unknown parameters and her joint priors on the bias and estimands, is the researcher better off conditioning her inferences solely on the study of her estimand of interest, using an evidence strategy of $\mathbf{y} = y_{\tau+\beta}$? Or can she improve her understanding of the world by using all of the evidence at hand, choosing $\mathbf{y} = [y_{\tau+\beta} \quad y_{\psi+\beta}]$ or even $\mathbf{y} = [y_{\tau+\beta} \quad y_{\psi+\beta} \quad y_{\psi}]$?

The following table represents all possible ways in which the data could be realized, along with two possible prior beliefs about the world.

	$\overbrace{\quad\quad\quad\quad}^{\Theta}$		$\overbrace{\quad\quad\quad}^{\theta}$			$\overbrace{\quad\quad\quad\quad\quad\quad}^{\mathcal{Y}}$		
e	$\Pr(\theta)_1$	$\Pr(\theta)_2$	τ	ψ	β	$y_{\tau+\beta}$	$y_{\psi+\beta}$	y_{ψ}
1	.125	.245	0	0	0	0	0	0
2	.125	.005	0	1	0	0	1	1
3	.125	.005	1	0	0	1	0	0
4	.125	.245	1	1	0	1	1	1
5	.125	.245	0	0	1	1	1	0
6	.125	.005	0	1	1	1	2	1
7	.125	.005	1	0	1	2	1	0
8	.125	.245	1	1	1	2	2	1

If the researcher has the prior belief $\Pr(\theta)_1$, she attributes equal probability to all e events, implying independence in the priors on estimands τ and ψ, $\rho_{\tau,\psi}(\Pr(\theta)_1) = 0$. By contrast, $\Pr(\theta)_2$ represents a belief that the unknown estimands are much more likely to share the same true value than different true values, such that $\rho_{\tau,\psi}(\Pr(\theta)_2) \approx 1$.

Denoting one realization of the observable evidence $\hat{\mathbf{y}}$, the posterior mean estimate of the unknown parameters given the evidence, $f(\theta \mid \hat{\mathbf{y}})$, is equal to

$$f\left(\theta|\,\hat{\mathbf{y}}, \Pr(\theta)\right) = \frac{\sum^{e} \Pr(\mathbf{y} = \hat{\mathbf{y}} \mid \theta = \theta_e) \Pr(\theta_e)\theta_e}{\Pr(\mathbf{y} = \hat{\mathbf{y}})} \tag{1}$$

$$= \frac{\sum^{e} \mathbf{1}(\mathbf{y}_e = \hat{\mathbf{y}}) \Pr(\theta_e)[\tau_e \quad \psi_e \quad \beta_e]}{\sum_{i:\mathbf{y}_i = \hat{\mathbf{y}}} \Pr(\theta_i)}. \tag{2}$$

By computing the posterior mean estimate of the parameters for each conceivable way in which the data can be realized, and weighting by the prior probability of each joint data event e occurring, we can compute the following loss function,

$$\mathcal{L}(\theta, \Pr(\theta), \mathbf{y}) = \mathrm{E}[(\theta - f(\theta \mid \Pr(\theta), \mathbf{y}))^2] \tag{3}$$

$$= \sum^{e} \Pr(\theta_e)(\theta_e - f(\theta \mid \Pr(\theta_e), \mathbf{y_e}))^2, \tag{4}$$

which is the expected squared error in the posterior estimates of the true unknown parameters, given all of the possible events that could give rise to the evidence, beliefs about the probability of those events

occurring, and the strategy used to synthesize evidence when forming posterior beliefs.

Definition 2. *The **quality of an inferential strategy** is defined by $\mathcal{L}(\theta, \Pr(\theta), \mathbf{y})$ given in equation 3. A study is of perfect inferential quality when $\mathcal{L}(\theta, \Pr(\theta), \mathbf{y}) = 0$, and decreases in quality as $\mathcal{L}(\theta, \Pr(\theta), \mathbf{y})$ increases.*

Conjecture 1. *Let $\beta_\tau = \psi_\tau$. Iff $\rho_{\tau,\psi}(\Pr(\theta)) < 1$ and $0 < \mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}])$, $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta} \quad y_{\psi+\beta}]) < \mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}])$.*

Conjecture 1 states that, if the joint prior does not induce perfect dependence in beliefs about estimands, the strategy of conditioning inferences on a single source of evidence is not perfectly informative, and the bias in the evidence on the primary estimand is the same as the bias in evidence on an ancillary estimand, conditioning inferences on both sources of evidence is provides strictly lower expected squared error in posterior inferences than conditioning only on the evidence about the primary estimand.

The possibility of this conjecture can be demonstrated using the table above. Suppose, for example, the researcher chooses to base her inferences solely on the evidence on her primary estimand, such that $\mathbf{y} = y_{\tau+\beta}$, and she observes $\hat{\mathbf{y}} = [y_{\tau+\beta} = 1]$. Then her posterior belief about the mean of the unknown parameters is

$$f(\theta \mid y_{\tau+\beta} = 1, \Pr(\theta)_1) = \frac{\sum^e \mathbf{1}(y_{\tau+\beta,e} = 1) \Pr(\theta_e)[\tau_e \quad \psi_e \quad \beta_e]}{\sum_{i:y_{\tau+\beta,i}=1} \Pr(\theta_i)} \tag{5}$$

$$\begin{aligned} &0 \times .125 \times [0 \quad 0 \quad 0] + 0 \times .125 \times [0 \quad 1 \quad 0] + \\ &1 \times .125 \times [1 \quad 0 \quad 0] + 1 \times .125 \times [1 \quad 1 \quad 0] + \\ &1 \times .125 \times [0 \quad 0 \quad 1] + 1 \times .125 \times [0 \quad 1 \quad 1] + \\ &= \frac{0 \times .125 \times [1 \quad 0 \quad 1] + 0 \times .125 \times [1 \quad 1 \quad 1]}{.125 + .125 + .125 + .125} \end{aligned} \tag{6}$$

$$= [.5 \quad .5 \quad .5] = [\tilde{\tau} \quad \tilde{\psi} \quad \tilde{\beta}]. \tag{7}$$

The intuition behind this estimate is clear: whereas observing a 0 or a 2 is perfectly informative about the value of the primary estimand and the bias, when the researcher observes a 1 she does not know whether it was generated due to bias or due to the estimand of interest because events 3-6 are observationally equivalent. For this reason, the posterior mean $\tilde{\theta} = [.5 \quad .5 \quad .5]$ is equal to the prior mean belief about the unknowns, $\sum^e Pr(\theta_e)\theta_e = [.5 \quad .5 \quad .5]$. In other words, even if the posterior variance shrinks the researcher learns nothing about the true location of the unknowns in this case.

Suppose, however, that she were to set her evidence synthesis strategy to $\mathbf{y} = [y_{\tau+\beta} \;\; y_{\psi+\beta}]$, and that she additionally observes $y_{\psi+\beta} = 0$. If she conditions her posterior beliefs on both pieces of observable evidence her strategy is completely informative about the true values of all unknown parameters:

$$f(\theta \mid y_{\tau+\beta} = 1, y_{\psi+\beta} = 0, \Pr(\theta)_1) = [1 \;\; 0 \;\; 0]. \tag{8}$$

The only way in which the evidence $\hat{\mathbf{y}} = [y_{\tau+\beta} \;\; y_{\psi+\beta}] = [1 \;\; 0]$ could have been observed is if the bias and ancillary estimand are both 0. Thus, simply by including a second study and exploiting a correlated bias assumption, the researcher has been able estimate the bias to be 0, and therefore increase her certainty about the quantity of interest, τ, despite having no prior knowledge about the location and scale of the bias or estimands.

In this simplified example, even though both sources of data are at risk of unknown bias about whose location and scale the researcher has flat priors, including the second kind of evidence renders the first kind informative. This finding is distinct from the banal claim that more data is better. In this case, for example, observing y_β twice implies $\mathbf{y} = [1 \;\; 1]$, which does not improve inferences. Rather, this example shows it is possible to estimate the bias and simultaneously update beliefs about the true estimand, because both sources of data provide independent information about the bias.

The following table provides the intuition about the independence condition in proposition 1:

			$\mathbf{y} = [y_{\tau+\beta}]$		$\mathbf{y} = [y_{\tau+\beta} \;\; y_{\psi+\beta}]$	
e	τ	ψ	$\Pr(\theta)_1(\tau - \tilde{\tau})^2$	$\Pr(\theta)_2(\tau - \tilde{\tau})^2$	$\Pr(\theta)_1(\tau - \tilde{\tau})^2$	$\Pr(\theta)_2(\tau - \tilde{\tau})^2$
1	0	0	0.000	0.000	0.000	0.000
2	1	0	0.031	0.001	0.000	0.000
3	0	1	0.000	0.000	0.000	0.000
4	1	1	0.031	0.061	0.031	0.061
5	0	0	0.031	0.061	0.031	0.061
6	1	0	0.000	0.000	0.000	0.000
7	0	1	0.031	0.001	0.000	0.000
8	1	1	0.000	0.000	0.000	0.000

Each row shows one possible state of the world, or data event. In the first, both the bias and the estimand are equal to 0, for example. The fourth and fifth columns show the squared error in the posterior that would result under that state of the world if one were to condition inferences only on the observation

of $y_{\tau+\beta}$ evidence, given independent and perfectly dependent priors on the estimands, respectively. The final two columns show the same information, albeit with inferences conditioned on both $y_{\tau+\beta}$ and $y_{\psi+\beta}$ data.

The greatest expected squared error in the posterior inferences occurs when the two studies produce the same results because their estimands are the same, and this occurs more often in expectation when the two estimands are correlated. Note that

$$\frac{\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta}])} < \frac{\mathcal{L}(\tau, \Pr(\theta)_2, [y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau, \Pr(\theta)_2, [y_{\tau+\beta}])} \tag{9}$$

$$.5 < .98. \tag{10}$$

The mean squared error with reference to the primary estimand is halved when conditioning on both studies under independent estimands: inferences about τ are better when the researcher considers a second source of data at risk of a common source of bias. However, if the estimands are highly correlated, the inferential gains from integration are much smaller.

Conjecture 2. *If* $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}]) > 0$ *and* $\rho_{\tau,\psi}(\Pr(\theta)) = 1$*, then* $\frac{\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}])} = 1$.

Proposition 2 is similar to the illusion of learning theorem: inferences are never improved through the integration of possibly biased data when that data does not provide independent information on the estimand or the bias. The intuition behind proposition 2 is made clear when one considers the case in which the data-generating processes of the evidence are exactly the same. If $y_{\tau+\beta} = y_{\psi+\beta}$, then $\frac{\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta}])} = \frac{\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta}])}{\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta}])} = 1$. This discussion makes clear why the principle of triangulation is useful.

Definition 3. *The bias inherent in a given source of evidence on an estimand is* ***triangulated*** *when it is estimated independently of that estimand through the use of two or more other pieces of evidence on an ancillary estimand.*

Suppose that the researcher sets her synthesis strategy to $\mathbf{y} = [y_{\tau+\beta} \quad y_{\psi+\beta} \quad y_{\psi}]$, so that she now conditions her inferences on an unbiased source of evidence about the ancillary estimand, ψ. Although she is not principally interested in ψ, she can use this ancillary estimand to obtain information on β without relying on her inferences about τ, because $\beta = y_{\psi+\beta} - y_{\psi}$.

Conjecture 3. *Iff* $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta} \quad y_{\psi+\beta}]) > 0$ *and* $\mathcal{L}(\beta, \Pr(\theta), [y_{\psi+\beta} \quad y_{\psi}]) < \mathcal{L}(\beta, \Pr(\theta), [y_{\psi+\beta}])$*, then* $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta} \quad y_{\psi+\beta} \quad y_{\psi}]) < \mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta} \quad y_{\psi+\beta}])$*, even if* $\mathcal{L}(\tau, \Pr(\theta), y_{\psi}) = \mathcal{L}(\tau, \Pr(\theta), \emptyset)$ *or* $\rho_{\tau,\psi}(\Pr(\theta)) = 1$.

Proposition 3 states that even if a source of evidence provides no independent information about the estimand of interest,[2] inferences about that same estimand are strictly improved by including it in an evidence synthesis strategy if it is informative about the correlated bias term. In fact, in the foregoing example, bias is perfectly estimated when unbiased evidence on ψ is included.

This is true for any level of dependence in the estimand priors. For example, if they are independent, $\mathcal{L}(\tau, \Pr(\theta)_1, [y_{\tau+\beta} \quad y_{\psi+\beta} \quad y_{\psi}]) = 0$. If the estimands are perfectly correlated in the prior, then y_{ψ} is perfectly informative about τ, since $y_{\psi} = y_{\tau}$. Thus, it is possible to use bias triangulation as an evidence synthesis strategy that improves inferences about an estimand, irrespective of the dependence in estimand priors.

2.2 Dependence of Biases

The exposition of the BEAMA strategy for data collection and analysis has so far assumed priors in which the biases are identical. Yet the logic generalizes to cases in which priors do not attribute perfectly joint distributions to the bias terms, but simply imply some prir dependence.

Assume that $\beta_{\tau} \neq \beta_{\psi}$, $y_{\tau+\beta} = \tau + \beta_{\tau}$ and $y_{\psi+\beta} = \psi + \beta_{\psi}$.

Conjecture 4. *Let* $\beta_{\tau} \neq \beta_{\psi}$. *If* $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}] > 0$, $\rho_{\beta_{\tau},\beta_{\psi}}(\Pr(\theta)) > 0$ *and* $\rho_{\tau,\psi}(\Pr(\theta)) < 1$, *then* $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta} \quad y_{\psi+\beta}]) < \mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}])$.

Proposition 4 states that when biases are not exactly the same across studies, inferences about an estimand are still improved through integration of the at-risk studies if estimand priors are not perfectly dependent and bias priors are not perfectly independent. Suppose now that the possible states of the

[2]Here $\mathcal{L}(\tau, \Pr(\theta), \emptyset)$ represents the expected squared error in the posterior when only prior information is used.

world are represented by the following table:

Θ		θ				$\mathcal{Y}$	
$\Pr(\theta)_3$	$\Pr(\theta)_4$	τ	ψ	β_τ	β_ψ	$y_{\tau+\beta}$	$y_{\psi+\beta}$
0.062	0.078	0	0	0	0	0	0
0.062	0.078	1	0	0	0	1	0
0.062	0.078	0	1	0	0	0	1
0.062	0.078	1	1	0	0	1	1
0.062	0.047	0	0	1	0	1	0
0.062	0.047	1	0	1	0	2	0
0.062	0.047	0	1	1	0	1	1
0.062	0.047	1	1	1	0	2	1
0.062	0.047	0	0	0	1	0	1
0.062	0.047	1	0	0	1	1	1
0.062	0.047	0	1	0	1	0	2
0.062	0.047	1	1	0	1	1	2
0.062	0.078	0	0	1	1	1	1
0.062	0.078	1	0	1	1	2	1
0.062	0.078	0	1	1	1	1	2
0.062	0.078	1	1	1	1	2	2

In this example, $\rho_{\beta_\tau,\beta_\psi}(\Pr(\theta)_3) = 0$: the researcher allows for any value of β_ψ given the true value of β_τ, and vice versa. In other words, she does not see the bias as similar across evidence sources in this specification of the joint bias and estimand priors. By contrast, under $\Pr(\theta)_4$ she allows for some correlation in her beliefs about the two bias terms, such that $\rho_{\beta_\tau,\beta_\psi}(\Pr(\theta)_4) = .5$. First, we can note that

$$\frac{\mathcal{L}(\tau, \Pr(\theta)_3, [y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau, \Pr(\theta)_3, [y_{\tau+\beta}])} = \frac{.125}{.125} = 1. \tag{11}$$

There are no inferential gains to including the second study when the biases are independent: the second study provides no new information on β_τ, and therefore does not improve inferences about τ. Inferences are improved if there is some correlation,

$$\frac{\mathcal{L}(\tau, \Pr(\theta)_4, [y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau, \Pr(\theta)_4, [y_{\tau+\beta}])} = \frac{.121}{.125} = .968, \tag{12}$$

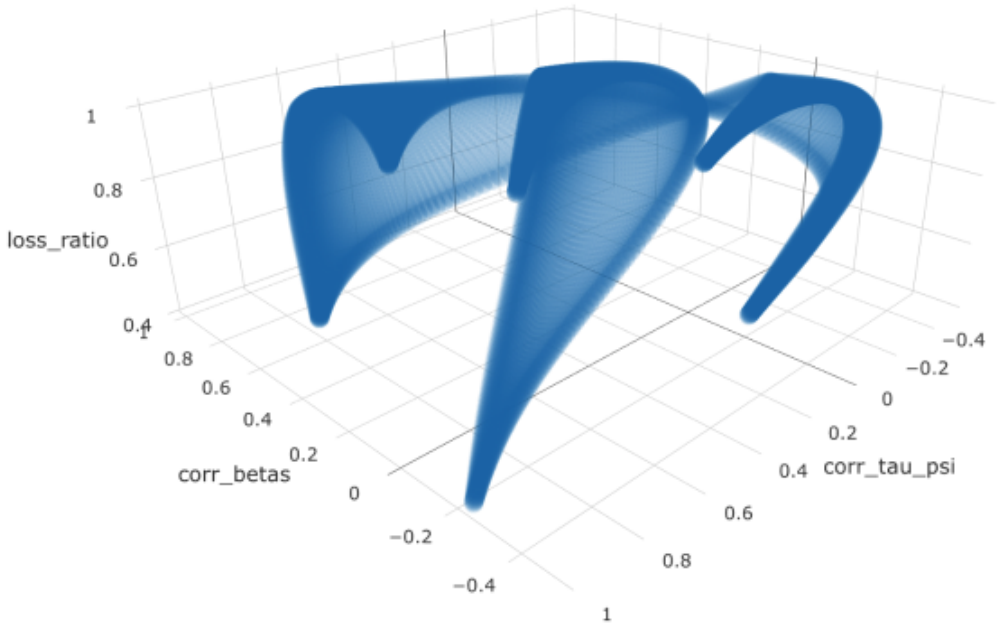

Figure 1: The ratio in expected mean squared error for synthetic and non-synthetic evidence strategies given flat marginal priors on the unknown parameters. The vertical axis shows $\frac{\mathcal{L}(\tau,\Pr(\theta),[y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau,\Pr(\theta),[y_{\tau+\beta}])}$, while the left and right horizontal axes display $\rho_{\beta_\tau,\beta_\psi}(\Pr(\theta))$ and $\rho_{\tau,\psi}(\Pr(\theta))$, respectively.

however the gain is minimal compared to the case in which the bias is perfectly correlated.

Conjecture 5. *If* $\mathcal{L}(\tau, \Pr(\theta), [y_{\tau+\beta}]) > 0$ *and* $\rho_{\beta_\tau,\beta_\psi}(\Pr(\theta)) = 0$*, then* $\frac{\mathcal{L}(\tau,\Pr(\theta),[y_{\tau+\beta} \quad y_{\psi+\beta}])}{\mathcal{L}(\tau,\Pr(\theta),[y_{\tau+\beta}])} = 1$.

Proposition 5 states that there are no inferential gains about the primary estimand from evidence synthesis if the bias priors exhibit no dependency across the sources of evidence used for triangulation. Intuitively, even when the ancillary bias is perfectly estimated, this is only helpful insofar as it is informative in some way about the possible values of the bias in the evidence on the primary estimand.

We can visualize the expected inferential gain of including a second study at risk of bias by integrating over a range of possible beliefs about bias and estimand correlation in Θ. Figure 1 plots the results of this exercise considering only those "flat priors" in which the marginal probabilities of the different values of unknowns are equal, i.e. in which the researcher knows nothing about the true location or scale of bias.

It can be seen that the loss ratio is lowest when priors on biases are perfectly correlated and priors on

estimands are negatively correlated. In fact, when the bias priors are correlated at ≈ 1 and the estimand priors at $\approx -.15$, the loss ratio drops to ≈ 0. Intuitively, this happens because ambiguous cases, in which $\hat{y}_{\tau+\beta} = 1$ and $\hat{y}_{\psi+\beta} = 0$, for example, are very unlikely to arise due to bias and very likely to arise due to a true effect. By contrast, if the priors on biases and estimands are weakly positively correlated, scenarios in which the bias or the estimands are non-zero are equally likely to produce such patterns and thus the evidence provides less probative value on average.

The foregoing discussion usefully demonstrates the possibility of bias estimation adjustment meta-analysis, using a setup that is simplified to ease exposition. In reality parameters will often be continuous and it is therefore necessary to learn about their location and scale. Furthermore, evidence is likely to be subject to multiple, potentially cross-cutting sources of systematic error, and it is less clear how well these kinds of bias can be estimated and how well doing so improves inferences. The following sections reiterate the core possibility results outlined above in applied statistical contexts, and examine further the conditions under which inferences are improved through bias estimation adjusted meta-analysis.

3 Common Sources of Bias

As section 2 demonstrated, the BEAMA approach differs from other approaches to bias-adjusted evidence synthesis insofar as it assumes similarity in bias across sources of evidence on different estimands. Consequently, it is essential to define the manner in which bias is arises within two sources of evidence, as beliefs about these mechanisms justify the use of priors in which beliefs about bias are correlated. In this section, I show that the mean and variance of error in a given evidence generation process can be decomposed according to features of the research design, and use this decomposition to formally define different sources of bias in an evidence generating process. These definitions highlight which kinds of meta-data are necessary to conduct BEAMA, clarify what kinds of assumptions about this data need to be defended, and inform the statistical model developed in the following section.

3.1 Identifying Bias-Generating Mechanisms Through Error Decomposition

The following presents an extension of Imai, King, and Stuart (2008), which decomposes the error of a difference-in-means estimator of a population average treatment effect.

Consider a large finite population of size N. A researcher generating evidence about some unknown quantity among the population conducts the m'th study on that quantity. I discuss causal estimands below (average differences among counterfactual states of the world), but the framework generalizes to

other quantities such as the true average of some feature of the sample. It is most natural perhaps to think of the study as a randomized controlled trial, but the setup applies to observational research and arguably to the qualitative evaluations of experts observing some causal process in the world.

Units indexed $i \in 1...N$ are sampled into the m'th study using the random variable $S_i \in \{0, 1\}$, and assigned to treatment according to the random variable $Z_i \in \{0, 1\}$, such that unit i is sampled when $S_i = 1$ (0 otherwise) and assigned to treatment when $Z_i = 1$ (0 otherwise). Let $V_m \in \{0, 1\}$ be an unobserved variable that varies at the study level, such that $V_m = 1$ if the study violates an exclusion restriction and 0 otherwise. Each unit has observed and unobserved covariates X_i, with λ_i denoting the effect on each unit's outcomes when exposed to the study-level exclusion restriction. The random variables are thus Z_i, S_i, X_i and λ_i, with V_m fixed for a given design. These variables are mapped to the potential outcomes of each unit through the following function

$$Y_i(Z, V) = Z_i g(X_i) + (1 - Z_i)h(X_i) + Z_i V_m \lambda_i, \tag{13}$$

where $g()$ and $h()$ are unknown functions. Ruling out exclusion restriction violations, the individual treatment effect is denoted

$$T_i = Y_i(Z = 1, V = 0) - Y_i(Z = 0, V = 0) \tag{14}$$

$$= g(X_i) - h(X_i) + 0 \times V_m \lambda_i \tag{15}$$

$$= g(X_i) - h(X_i). \tag{16}$$

The main estimand of interest is the population average treatment effect (PATE):

$$\tau = \frac{1}{N} \sum_{i=1}^{N} T_i. \tag{17}$$

Let $N(S) = |\{i \ : \ S_i = 1\}|$ (the number of sampled units) and $N(Z) = |\{i \ : \ S_i = 1, Z_i = 1\}|$ (the number of assigned units, conditional on sampling). Define two more estimands that are not of primary interest. The first is the sample average treatment effect (SATE):

$$\tau^S = \frac{1}{N(S)} \sum_{i:S_i=1}^{N(S)} T_i, \tag{18}$$

and the second is the average difference in potential outcomes between those in the sample assigned to

treatment or control, minus the exclusion restriction:

$$\tau^Z = \frac{1}{N(Z)} \sum_{i:S_i=1,Z_i=1}^{N(Z)} (Y_i - V_m\lambda) - \frac{1}{(N(S) - N(Z))} \sum_{i:S_i=1,Z_i=0}^{N(S)-N(Z)} Y_i \tag{19}$$

$$= \frac{1}{N(Z)} \sum_{i:S_i=1,Z_i=1}^{N(Z)} (g(X_i)) - \frac{1}{(N(S) - N(Z))} \sum_{i:S_i=1,Z_i=0}^{N(S)-N(Z)} h(X_i). \tag{20}$$

And finally the estimate is equal to :

$$\hat{\tau} = \frac{1}{N(Z)} \sum_{i:S_i=1,Z_i=1}^{N(Z)} Y_i - \frac{1}{(N(S) - N(Z))} \sum_{i:S_i=1,Z_i=0}^{N(S)-N(Z)} Y_i \tag{21}$$

Note that τ, the true average of the individual causal effects among the population, and τ^S, the true average of the individual causal effects among the sample, are both fundamentally unobservable because Z_i cannot be 0 and 1 at any given time. In principle, however, τ^Z is observable, since $\hat{\tau}$ is observable and for studies without exclusion restriction violations $\hat{\tau} = \tau^Z + V_m\lambda_i = \tau^Z$.

Using these four definitions, we can decompose the error of an estimator, $\hat{\tau} - \tau$, into three sources of error based on the research design:

$$\epsilon^S = \tau^S - \tau, \qquad \epsilon^Z = \tau^Z - \tau^S, \qquad \epsilon^V = \hat{\tau} - \tau^Z. \tag{22}$$

The first term denotes error arising from the sampling strategy, the second denotes error arising from the assignment strategy, and the third term denotes error arising from a violation of the exclusion restriction. It is worth noting here that other sources of error are possible given more complex research designs. For example, this setup contains a stable unit treatment value assumption (SUTVA), and does not account for additional sources of error that may arise from the estimation strategy. Moreover, attrition and non-compliance could intervene at various stages of a design, biasing estimates. However, this framework is sufficiently general to capture a wide variety of evidence-generating processes used in observational, experimental and (arguably) even qualitative research.

Rearranging the terms in equation 22, the evidence-generating process can thus be written:

$$\hat{\tau} = \tau + \epsilon^S + \epsilon^Z + \epsilon^V. \tag{23}$$

Since τ is constant (see equation 17), we can write the expected value of the e'th source of evidence as

follows,

$$\mathrm{E}[\hat{\tau}] = \tau + \mathrm{E}[\epsilon^S] + \mathrm{E}[\epsilon^Z] + \mathrm{E}[\epsilon^V]. \tag{24}$$

This provides the following definitions.

Definition 4. *An* unbiased *source of evidence on an estimand τ is one in which $\mathrm{E}[\hat{\tau}] = \tau$, since $\mathrm{E}[\hat{\tau} - \tau] = 0$.*

Definition 5. *A study is at risk of* sampling bias *if $\mathrm{E}[\hat{\tau} - \tau - \epsilon^Z - \epsilon^V] = \mathrm{E}[\epsilon^S] \neq 0$, where the amount of sampling bias is equal to $\mathrm{E}[\epsilon^S]$.*

Definition 6. *A study is at risk of* assignment bias *if $\mathrm{E}[\hat{\tau} - \tau - \epsilon^S - \epsilon^V] = \mathrm{E}[\epsilon^Z] \neq 0$, where the amount of assignment bias is equal to $\mathrm{E}[\epsilon^Z]$.*

Definition 7. *A study is at risk of* exclusion restriction bias *if $\mathrm{E}[\hat{\tau} - \tau - \epsilon^S - \epsilon^Z] = \mathrm{E}[\epsilon^V] \neq 0$, where the amount of exclusion restriction bias is equal to $\mathrm{E}[\epsilon^V]$.*

Noting $\mathrm{Var}[\tau] = 0$ and denoting Σ the variance-covariance matrix of the errors defined in 22, the variance of an evidence generating process is equal to

$$\begin{aligned} \mathrm{Var}[\hat{\tau}] = & \quad \mathrm{Var}[\epsilon^S] + \mathrm{Var}[\epsilon^Z] + \mathrm{Var}[\epsilon^V] + \\ & 2(\mathrm{Cov}(\epsilon^S, \epsilon^Z) + \mathrm{Cov}(\epsilon^V, \epsilon^Z) + \mathrm{Cov}(\epsilon^S, \epsilon^V)) \end{aligned} \tag{25}$$

$$= \mathbf{1}'\Sigma\mathbf{1}. \tag{26}$$

We can then assume that these different sources of bias have distributions. For example, if a researcher has evidence on τ that she suspects is at risk of sampling bias, $y_{\tau+\beta^S}$, but knows it is free of assignment bias and is virtually certain that it is absent of exclusion restriction bias, she might assume

$$y_{\tau+\beta^S} \sim \mathcal{N}(\tau + \epsilon^S + \epsilon^Z + \epsilon^V, \Sigma) \tag{27}$$

$$\epsilon^S \sim \mathcal{N}(\beta^S, \sigma^S) \tag{28}$$

$$\epsilon^Z \sim \mathcal{N}(0, \sigma^Z) \tag{29}$$

$$\epsilon^V \sim \mathcal{N}(0, .001). \tag{30}$$

The framework therefore allows for informative or uninformative priors about bias, and so generalizes to cases in which researchers wish to use the opinions of experts in their estimation strategy. In general,

however, I assume researchers do not know the location of unknown bias, and rarely know the scale.

The next section extends upon these models to account for multiple studies with varying risk of bias. Before discussing models that leverage knowledge about bias and bias decomposition across sources of evidence, however, it is useful to demonstrate some of the procedures outlined above and to discuss which kinds of assumptions seem defensible in practice, as these help to inform the modeling strategy.

3.2 An Illustration

To illustrate the concept of error decomposition, consider two research designs aiming at the same estimand and focusing on the same population. Let p_1^S and p_2^S denote the sampling probabilities of designs 1 and 2, and and p_1^Z and p_2^Z their (unconditional) assignment probabilities. Assume that these probabilities are not known to the researcher and so are not used as data in the estimation strategy. For an imagined implementation of the study, vectors $\hat{S}_1, \hat{S}_2$ denote realizations of the sampling procedure, $\hat{Z}_1, \hat{Z}_2$ denote realizations of the assignment procedure, and $\hat{Y}_1, \hat{Y}_2$ denote realizations of the outcomes. All values are rounded to the first decimal place. Table 1 shows one possible realization of the study under the two designs. The first four columns are labeled with the potential outcomes, $Y_i(Z, V)$ as defined in equation 13.

In table 1, the true PATE estimand is $\tau = 1.02$, while the first estimate is $\hat{\tau}_1 = 2.12$ and the second is $\hat{\tau}_2 = 2.18$. It is straightforward to verify that $\hat{\tau}_1 = \tau + \epsilon_1^S + \epsilon_1^Z + \epsilon_1^V = 1.02 + 0.32 + 0.59 + 0.19 = 2.12$ and $\hat{\tau}_2 = \tau + \epsilon_2^S + \epsilon_2^Z + \epsilon_2^V = 1.02 + .29 + .67 + .20 = 2.18$. More interesting are the distributions of these errors, the densities of which are plotted on figure 2.

Note firstly that the only difference in these designs are the probabilities with which units are sampled and assigned. Both set $N(S) = 6$ and $N(Z) = 3$, however the first uses a sampling strategy that selects units with equal probability while assigning treatment with heterogeneous probabilities, while the second design assigns treatment with equal probabilities but samples using heterogeneous probabilities. The first

$Y(1,0)$	$Y(0,0)$	$Y(1,1)$	$Y(0,1)$	p_1^S	p_1^Z	p_2^S	p_2^Z	$\hat{S}_1$	$\hat{Z}_1$	$\hat{Y}_1$	$\hat{S}_2$	$\hat{Z}_2$	$\hat{Y}_2$
1.1	−0.6	1.3	−0.6	0.1	0.0	0.2	0.1	1	0	−0.6	0	–	–
1.8	0.2	2.0	0.2	0.1	0.2	0.2	0.1	0	–	–	1	1	2
0.7	−0.8	0.9	−0.8	0.1	0.0	0.2	0.1	1	0	−0.8	1	0	−0.8
2.9	1.6	3.1	1.6	0.1	0.2	0.2	0.1	1	1	3.1	1	1	3.1
2.0	0.3	2.2	0.3	0.1	0.0	0.2	0.1	1	0	0.3	1	0	0.3
0.7	−0.8	0.9	−0.8	0.1	0.2	0.2	0.1	1	1	0.9	1	0	−0.8
0.6	0.5	0.8	0.5	0.1	0.0	0.0	0.1	0	–	–	0	–	–
1.0	0.7	1.2	0.7	0.1	0.2	0.0	0.1	1	1	1.2	0	–	–
0.8	0.6	1.0	0.6	0.1	0.0	0.0	0.1	0	–	–	0	–	–
−0.1	−0.3	0.1	−0.3	0.1	0.2	0.0	0.1	0	–	–	1	1	0.1

Table 1: One realization of a single study under two different designs.

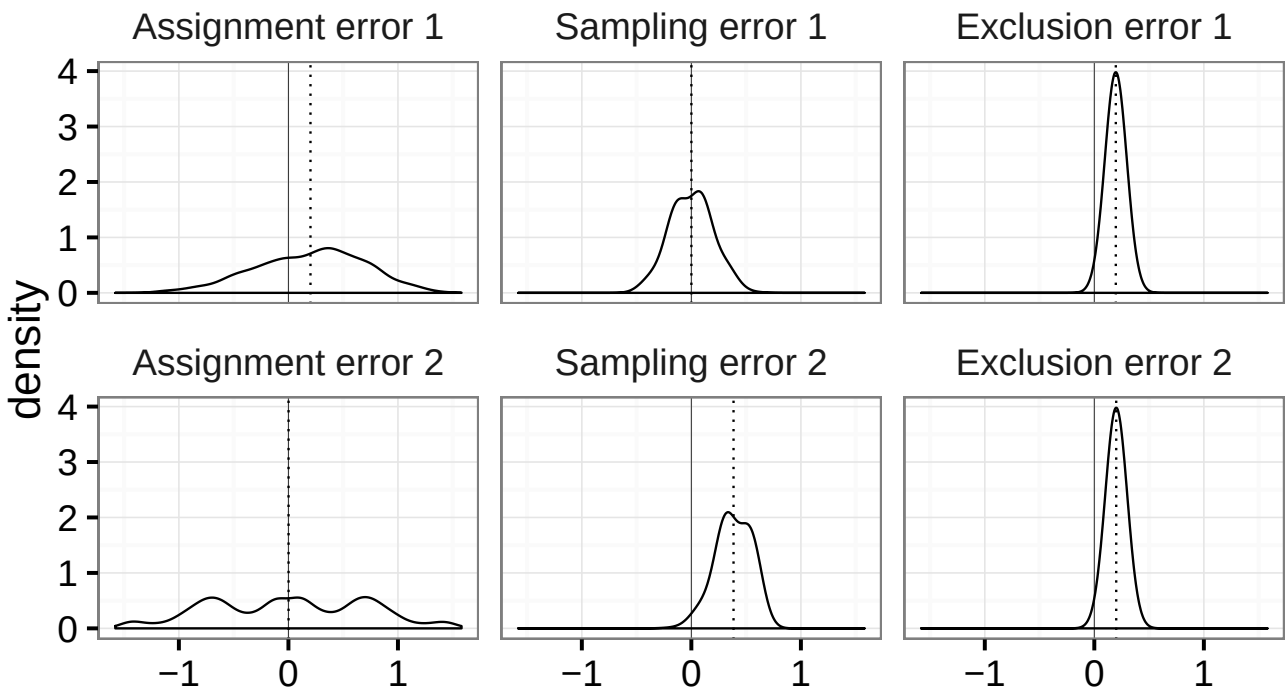

Figure 2: Distribution of decomposed errors of a single study under two alternative designs.

design is thus analogous to an observational study in which units are selected from the population at random, while the second is akin to an RCT that uses a convenience sample.

Both designs produce biased estimates. The first design's assignment procedure produces bias equal to $\mathrm{E}[\epsilon_1^Z] = .20$, but its sampling procedure is unbiased $\mathrm{E}[\epsilon_1^S] = 0$. The converse is true for the second design, which features an unbiased assignment procedure $\mathrm{E}[\epsilon_2^Z] = 0$, and a biased sampling procedure $\mathrm{E}[\epsilon_2^S] = .38$. Thus, if these designs represented evidence on the same estimand arising from two separate studies, the researcher would be justified in treating their sampling and assignment bias parameters as separate.

However, the designs share an identical exclusion restriction bias, where $\mathrm{E}[\epsilon_1^V] = \mathrm{E}[\epsilon_2^V] = .20$, and $\mathrm{Var}[\epsilon_1^V] = \mathrm{Var}[\epsilon_2^V] = .00002$. In this case, the true effect of the exclusion restriction violation is akin to an estimand that a researcher treats as common across two sources of evidence, and doing so should be justified in a similar way. For example, the thing thought to give rise to the exclusion restriction violation – such as non-blinding, sponsorship bias, or another feature of treatment implementation that varies with the assignment – should be the same across studies. In this case, the researcher would be justified in assuming that the two studies share a common exclusion restriction bias.

One feature of this analysis that is also important to note is the bias in the frequentist estimates

of the variance. If we use Neyman standard errors to approximate the variance in the estimator, we obtain $\mathrm{E}[\widehat{\mathrm{Var}[\hat{\tau}]}] = .50$, whereas the true variance in the sampling distribution of the estimates is much lower, $\mathrm{Var}[\hat{\tau}] = .31$. This is because standard methods for computing the variance in the difference in means estimator in finite samples tend to underestimate the covariance in the potential outcomes (Splawa-Neyman et al., 1990; Aronow et al., 2014). This matters in section 4, because it implies that the Bayesian approach that incorporates these variance estimates and inflates the variance proportionally to the uncertainty in the prior is at very little risk of over-stating the certainty about any given estimate from a study.

3.3 Justifying Correlated Priors on Bias

The foregoing discussion illustrates that it is possible to decompose the error in a given design strategy with respect to an estimand, given our knowledge of that design strategy. However, it also highlights that even simple designs can give rise to multiple, cross-cutting forms of bias.

As outlined in section 2, BEAMA is likely to be most effective when the bias in two or more sources of evidence works in a similar way, such that estimating the bias in one source of evidence is informative about the scale and location of the bias in another source of evidence. It is therefore worth briefly discussing the kinds of situations in which it makes sense to assume that bias will manifest in like ways across evidence on different estimands.

With respect to sampling bias, it is most likely that two sources of evidence exhibit similar systematic sampling error when they use the same strategy on the same population. An extreme case of this might involve a single study. Suppose, for example, that a researcher is interested in τ and plans to conduct a randomized controlled trial to estimate this estimand, but for budgetary or practical reasons she is forced to conduct the experiment on a convenience sample. Suppose that she has evidence on ψ that comes from a randomized controlled trial conducted on a random sample from the same population in which she is interested. By incorporating a treatment into her own experiment that yielded information on ψ at the same time as it yielded information on τ, she would obtain $y_{\tau+\beta^S}$ and $y_{\psi+\beta^S}$ from her experiment. Using the available evidence on the ancillary estimand, y_ψ and $y_{\psi+\beta^S}$, she could triangulate β^S and thereby improve inferences about τ.

A similar logic applies to the integration of observational studies with RCTs. Suppose that there is a specific population of patients for which a range of observational and experimental sources of evidence on ψ exists. For ethical reasons, suppose that there are only very few or no randomized studies on the effects of τ in this population. There may be reason to believe that the allocation method employed is

similar across the non-randomized studies. By combining the evidence on randomized and non-randomized studies, the researcher can form reasonable beliefs about β^Z independently of τ, and thus improve her inferences about her primary estimand of interest.

There are many applications in which it is plausible to assume similar exclusion restriction violations may appear across sources of evidence, and possibly even across populations. For example, non-blinding of treatment conditions is frequently pointed to as a source of bias in placebo-controlled RCTs. If most of the studies that exist on τ were non-blinded, it should be possible to benchmark β^V using similar studies in which there was blinding and non-blinding, in order to estimate β^V independently from τ. If the non-blinding effect is large and positive across a range of outcomes, the researcher is better justified in assuming that it will also be large in a subsequent piece of evidence for which no blind evidence exists.

Another line of defense resides in the estimated heterogeneity of the bias. Suppose a researcher were to use studies on several ancillary estimands to triangulate the effect of sponsorship bias, and found that the standardized effect was very homogeneous. This lends empirical support to the assertion that this form of bias will operate similarly in the evidence on the primary estimand of interest, without ever relying on subjective beliefs about the nature of that bias.

One difficulty not addressed above is the question of the scale and value with which outcomes or estimates are encoded. On the one hand, different sources of evidence may encode information on the same process differently. For example, one source of evidence may report outcomes on an ordinal scale, while another reports them on a normal scale. In such cases, latent variable models may help to bring the potential outcomes into a common scale. Less problematically, estimates may be similarly valued but scaled in different ways. In such cases it should be feasible to standardize estimates accordingly in order to bring them to a common scale.

4 Statistical Approach

In applied meta-analysis settings, researchers rarely have data structured as the sum of binary components as the example in section 2 assumed. Rather, evidence on some estimand is typically represented by a point estimate of the estimand of interest, as well as an estimate of the uncertainty around that point estimate. Thus, to aggregate inferences of this kind into a joint posterior distribution, we need a statistical model that is able to take advantage of these two kinds of information, and that appropriately propagates uncertainty.

4.1 The BEAMA Model

In section 2, θ represented all unknowns. Now let θ represent the vector of unknown estimands, indexed by t. A meta-analysis may focus on many different estimands. For ease of exposition and in continuity with the foregoing examples, assume that there are two estimands so that $\theta = [\tau \quad \psi]$, and $t \in \{1, 2\} \quad : \quad \theta_1 = \tau, \quad \theta_2 = \psi$. Whereas before a single source of evidence yielded a scalar, now a source of evidence yields a vector, $y_m = [\hat{\theta}_{tm} \quad \hat{\sigma}^2_{\theta_{tm}}]$. For example, the m'th study of τ would yield a point estimate, $\hat{\tau}_m$, and an estimate of the variance around that point estimate, $\hat{\sigma}^2_{\tau_m}$. In the following, assume that these estimates come from non-Bayesian models, i.e. models that do not treat θ_t as a random variable in the estimation. Therefore, this data is an estimate of the quantity defined in equation 25, $\hat{\sigma}^2_{\tau_m} = \widehat{\text{Var}[\hat{\tau}_m]}$.

In addition to grouping the m'th piece of evidence according to the estimand it seeks to estimate, it is also possible to use the bias decomposition defined in the preceding section to sort the evidence into non-nested groups, according to our knowledge of the data-generating process used in each source of evidence. Suppose, for example, that two studies use identical sampling methods and different assignment methods. Suppose, further, that the first study estimates both τ and ψ. In this case, we have three pieces of evidence: one on τ, two on ψ. We know that the estimates of τ and ψ that come from the same study have very similar sampling and assignment errors. We know that the assignment errors should be treated differently between the two studies, but we may want to treat the sampling errors as drawn from the same distribution across all three sources of evidence, if the mechanisms are similar enough.

In the model, sampling groups are indexed $j \in \{1...J\}$, assignment groups $k \in \{1...K\}$, and exclusion groups $l \in \{1...L\}$. We can thus put the single-study likelihood defined in equation 27 into the meta-analysis context as follows,

$$\hat{\theta}_{tm} \sim \mathcal{N}(\theta_t + \epsilon^S_{jm} + \epsilon^Z_{km} + \epsilon^V_{lm}, \tilde{\sigma}_m). \qquad \tilde{\sigma_m} = \sqrt{(\hat{\sigma}^2_{\theta_{tm}} + \sigma^2_t)}. \tag{31}$$

The first equation states that the point estimate of the t'th estimand from the m'th study is distributed normally, with a mean equal to the sum of the true value of the t'th estimand of that study and all of the errors arising from the particular study design. The standard deviation of the m'th study's estimates of the estimand is defined as a partially-known quantity. Note that even though the estimand is defined in this framework as a constant (the PATE), it is an unknown constant about which we have beliefs and thus it is necessary to assign to it a prior probability density.

Thus, the standard deviation of the m'th point estimate is equal to the squared estimated standard

error inflated by the variance in our beliefs about the estimand. This follows directly from the definitions in section 3. Rather than treating the estimated variance of an estimate as fully known, however, this specification places a soft constraint on the variance in a given study. Referring to equations 25 and 26, the sum of the posterior estimates of the variance and covariance in the errors of a given study can never be less than the estimated variance of the estimator, $\mathbf{1}'\widetilde{\mathbf{\Sigma}}\mathbf{1} \geq \hat{\sigma}^2_{\tau_m}$, even if posterior uncertainty about θ_{tm} is reduced to 0. Thus, by inflating uncertainty about the variance in a given study through the addition of prior uncertainty about the estimand, we allow for situations in which a given study under-estimates the variance of its estimator. This might happen, for example, if 10 studies belong to exactly the same error group, and one of the studies produced an abnormally low variance estimate. In this case, the posterior inflation of the variance on that study would be incorporated into the posterior distribution over the estimand.

Priors on estimands are expressed as follows,

$$\theta_t \sim \mathcal{N}(\mu_t, \sigma_t), \qquad \mu_t \sim \mathcal{N}(0, 10), \qquad \sigma_t \sim \text{Half-Cauchy}(0, .5). \tag{32}$$

The estimand is a constant about which we have normally-distributed beliefs. The prior on the mean is weakly informative and implies diffuse beliefs about the true location of the estimand. Following Gelman et al. (2006), I put half-cauchy priors on all variance components. Here, the scale parameter is set to .5, which reflects the knowledge that the true variance in the estimand is 0. However, while much of the probability mass is thus concentrated near 0, this specification puts positive probability mass along the entire positive tail of the distribution. Thus, the posterior estimate of the variance in a given study can be inflated to arbitrarily large values provided noisy data that under-estimates uncertainty at the study level.

In accordance with section 3, the errors are modeled as partially known given our knowledge of the evidence generation process in a given study. Specifically, I make use of the decomposition defined in equations 25 and 26. The error terms in a given study are distributed multivariate-normally (MVN) as follows,

$$\boldsymbol{\epsilon} \sim \text{MVN}(\boldsymbol{\beta}, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = \text{diag}(\boldsymbol{\sigma})\mathbf{\Omega}\text{diag}(\boldsymbol{\sigma}), \tag{33}$$

where $\mathbf{\Sigma}$ is the variance covariance matrix of the errors, $\boldsymbol{\sigma}$ is a vector of error standard deviations, and $\mathbf{\Omega}$ is a correlation matrix.[3] Note that this specification involves an assumption of independence between

[3]The symbol Ω was also used in section 2 to denote the beliefs of the researcher about the possible states of the world that

the estimand and the errors. This is necessary here given equation 24: the correlation between a constant and a random variable is undefined due to the 0 in the denominator. This definition also justifies the independence in the priors on the estimands.

One possible arrangement of these errors for a meta-analysis with 5 pieces of evidence is represented in equations 34 and 35. Equation 35 represents possible priors on the unknown errors.

Turning firstly to the definitions in 34, note that the vector of errors is of length $5 \times J + 5 \times K + 5 \times L$, whereas the vector of error means (biases) is of length $J + K + L$. That is, a piece of evidence is not assumed to have literally the same error as other pieces of evidence in the same error group: rather, those studies have error drawn from the same distribution. To visualize these distributions, see figure 2: two studies with error that is 0 *on average* will nevertheless have some non-zero error *in a given realization*, depending on the distribution of that error.

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^S_{jm} \\ \vdots \\ \epsilon^S_{J5} \\ \epsilon^Z_{km} \\ \vdots \\ \epsilon^Z_{K5} \\ \epsilon^V_{lm} \\ \vdots \\ \epsilon^V_{L5} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta^S_j \\ 0 \\ \vdots \\ \beta^S_J \\ \beta^Z_k \\ \beta^Z_{k+1} \\ \vdots \\ \beta^Z_K \\ 0 \\ \beta^V_{l+1} \\ \vdots \\ \beta^V_L \end{bmatrix}, \qquad \boldsymbol{\sigma} = \begin{bmatrix} \sigma^S_j \\ \sigma^S_{j+1} \\ \vdots \\ \sigma^S_J \\ \sigma^Z_k \\ \sigma^Z_{k+1} \\ \vdots \\ \sigma^Z_K \\ \sigma^V_l \\ \sigma^V_{l+1} \\ \vdots \\ \sigma^V_L \end{bmatrix}, \qquad \boldsymbol{\Omega} = \begin{bmatrix} 1 & \cdots & \rho^{SV}_{J,L} \\ \vdots & \ddots & \vdots \\ \rho^{SV}_{J,L} & \cdots & 1 \end{bmatrix}. \tag{34}$$

$$\begin{aligned} \beta^S_j &\sim \mathcal{N}(0, 10) & \sigma^S_j &\sim \text{Half-Cauchy}(0, 1) \\ \beta^S_{j+1} &= 0 & \sigma^S_{j+1} &\sim \text{Half-Cauchy}(0, 1) \\ \beta^S_J &\sim \mathcal{N}(0, 10) & \sigma^S_J &\sim \text{Half-Cauchy}(0, 1) \\ &\vdots & &\vdots \\ \beta^V_L &\sim \mathcal{N}(0, 10) & \sigma^V_L &\sim \text{Half-Cauchy}(0, .001) \end{aligned} \qquad \boldsymbol{\Omega} \sim \text{LKJ}(2). \tag{35}$$

could obtain. Note that these were also beliefs about the correlation of estimands and bias. Thus, I retain the symbol here to maintain parallelism with the motivating example, and point out that the model essentially specifies weakly informative priors about correlation in bias and estimands.

Note that the vector $\boldsymbol{\beta}$ contains both unknown parameters and constants. This represents the partial knowledge of the researcher: whereas she knows that the sampling error of studies belonging to group $j = 2$ is 0 on average, she does not know what the average sampling error is for studies belonging to group $j = 1$. This is reflected in the column of equations below $\boldsymbol{\beta}$ in line 35: the priors are diffuse for some bias terms and set to known as 0 for others. This is the crucial mechanism in the model that allows for the bias triangulation procedure defined in section 2.

The vector of error standard deviations $\boldsymbol{\sigma}$ is also of length $J + K + L$. All true standard deviations are assumed unknown. However, as the priors on the standard deviations illustrate, the researcher might believe that certain kinds of bias do not vary greatly within groups. For example, if the researcher thinks of exclusion restriction violations as a population-level estimand (i.e. the effect of non-blinding in a population), this implies that the bias is constant, and does not contribute greatly to the overall variance in a given study.

The correlation in errors is assumed unknown. Following Lewandowski, Kurowicka, and Joe (2009) and Carpenter et al. (2016), I use the LKJ (Lewandowski-Kurowicka-Joe) correlation density to express a diffuse prior on the error correlations that is guaranteed to generate positive-definite variance-covariance matrices. The LKJ density has a single hyperparameter, η. For $\eta = 1$ the density is uniform over all possible correlation matrices, as η increases above 1 progressively more probability mass is concentrated on the identity matrix. Thus, the prior of $\eta = 2$ represents a weakly informative belief that the error correlations lie somewhere between -1 and 1.

Denoting the vector of unknown bias parameters $\boldsymbol{\beta}^{\text{un}}$, the vector of known bias parameters $\boldsymbol{\beta}^{\text{kn}}$, and the vector of estimands $\boldsymbol{\theta}$, the joint posterior distribution of the BEAMA model can thus be denoted

$$f(\boldsymbol{\theta}, \boldsymbol{\epsilon}, \boldsymbol{\beta}^{\text{un}}, \boldsymbol{\Omega}, \boldsymbol{\sigma} \mid \mathbf{y}, \boldsymbol{\beta}^{\text{kn}}) \propto \begin{array}{c} f(\boldsymbol{\theta}) f(\boldsymbol{\beta}^{\text{un}}) f(\boldsymbol{\sigma}) f(\boldsymbol{\Omega}) \times \\ f(\boldsymbol{\epsilon} \mid \boldsymbol{\Omega}, \boldsymbol{\sigma}, \boldsymbol{\beta}^{\text{un+kn}}) f(\mathbf{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}). \end{array} \tag{36}$$

4.2 Validation and Convergence

To fit the model, I program the above in Stan (Carpenter et al., 2016), which samples from the posterior distribution using Markov Chain Monte Carlo (MCMC) implemented through the No-U-Turn (NUTS) sampler (Hoffman and Gelman, 2014). NUTS employs an adaptive random walk algorithm based on Hamiltonian dynamics, and has been shown to quickly converge to target distributions with very little correlation in samples.

A minimal requirement of the model would be that it performs accurately in the simplest of settings.

To illustrate that this is so, I fit the model to a contrived data set. Specifically, there are 15 studies of three kinds: 5 that estimate τ with unknown bias β, 5 that estimate ψ with the same unknown bias β, and 5 that estimate ψ with no bias. Assume that there is no variance in estimates, and no variance in the bias terms. Setting $\tau = 1$, $\psi = -1$ and $\beta = .2$, this gives five datapoints equal to 1.2, five equal to -.8, and five equal to -1, with all 15 variance estimates equal to 0.

Using this simple setup, the model produces results identical to the posterior distributions described in section 2: the data is perfectly informative. In other words, the posterior distribution provided by the statistical model fit in Stan gives $\tau = 1, \psi = -1, \beta = .2$, with no variance in the posterior. The model is thus minimally valid insofar as it generalizes to the motivating examples. In the next section I assess how it performs in more challenging, realistic settings.

Another important consideration given the use of MCMC to approximate the posterior distribution is the extent to which samples converge on the true joint posterior distribution. If sampling converges to the wrong distribution, this will bias estimates. In principle this is not a strong concern given that in simulations the true values of all of the parameters are known, but ensuring a convergent model enables us to distinguish MCMC error from error in the inferential strategy employed by BEAMA.

The appendix contains plots of convergence diagnostics. Figure 9 shows the trace plots for the estimand, bias and error parameters. A convergent model is one that illustrates good 'mixing', i.e. one in which the parameter values converge to some stationary equilibrium.

The trace plots show occasional divergences for the error parameters, likely due to the stringent constraints on the variances in the model. Future work will seek to refine the computational aspects of the model to sample more efficiently. On average, however, all chains exhibit good mixing.

A more formal examination of the convergence properties involves the correlation among samples from the posterior within the 'chains' of samples. If the model has converged to the true posterior, draws of a single parameter should be uncorrelated with themselves within and across chains. Figure 10 illustrates that the autocorrelation in chains drops to approximately 0 after 30 draws. I thus proceed under the assumption that the implementation in Stan approximates the true posterior distribution of the BEAMA model.

5 Simulation Study

The simulation analysis aims to explore the inferential potential of BEAMA in applied meta-analysis settings. In all simulations, I use the package `DeclareDesign` to generate simulated studies that implement

all steps of a research design as a researcher would in the real world. The potential outcomes of a large finite population are defined, a sample is drawn, an explanatory variable (treatment) is allocated among the sample with some probability density, and the estimand is estimated. This process can be repeated under different designs to simulate different evidence-generating processes.

This approach has the dual advantage of creating known distributions that can be compared to the posterior estimates from BEAMA, while remaining as true as possible to the way in which evidence is actually generated. The code used to generate the examples is contained in the appendix.

I conduct four simulation experiments that analyze different aspects of BEAMA.

- In the **three-study meta-analysis with one unknown bias**, I simulate the inferential tradeoffs of different evidence-synthesis strategies that a researcher faces when she has one piece of evidence on her estimand of interest that is at risk of bias, one piece of evidence on an ancillary estimand that is at risk of a common bias, and one unbiased piece of evidence on the ancillary estimand.

- In the **medium-sized meta-analysis with one unknown bias**, I retain the same three kinds of evidence as used in the first simulation, but vary the different proportions of evidence used in the evidence strategy, from a minimum of one each to a maximum of 10 each (30-study meta-analysis).

- In the **medium-sized meta-analysis with incorrectly specified bias**, I assess how BEAMA performs in a context similar to the medium-sized analysis, but in which two different biases are wrongly specified as common.

- In the **medium-sized meta-analysis with two unknown biases**, I assess how BEAMA performs in a context in which there are multiple kinds of bias and estimands, and show the marginal improvement in inferences from including a study that enables triangulation of one kind of bias.

5.1 Three-Study Meta-Analysis with One Unknown Bias

In the motivating examples used in section 2, I showed that it is possible to greatly improve posterior inferences about an estimand, even when the only piece of information on that estimand is at risk of unknown bias. However, the data-generating process was highly unrealistic.

In this simulation study, I use `DeclareDesign` to generate three hypothetical sources of evidence that are analogous to the motivating example, albeit with a realistic data-generating process that lets bias arise as a result of the design.

The samples used to generate the three pieces of evidence are drawn from the same finite population of size $N = 100,000$. They use an identical procedure in which 2000 individuals are sampled at random with equal probability. The first piece of evidence estimates the estimand τ, while the second and third aim at the estimand ψ. The first two pieces of evidence, y_1 and y_2, are generated through a research design that assigns treatment to 250 units using uneven probabilities: moreover, those that are most likely to receive treatment also have higher potential outcomes. The last piece of evidence, y_3, is also generated through a design that assigns treatment to 250 units in the sample, but it does so using equal assignment probabilities. None of the studies violates the exclusion restriction.

The three pieces of evidence mimic different research situations. For example, one might consider might imagine a researcher confronted by two published studies, one observational and one experimental. The observational study contains estimates of the two different estimands, generating evidence y_1 and y_2, but the researcher does not know the assignment probabilities and so cannot know whether the estimates are biased. However, she knows that the estimate of the ancillary estimand in the experimental study, y_3, is unbiased. She is reasonably confident that the exclusion restriction is met in all of the pieces of evidence, but allows for some small amount of variation around 0 in this source of error.

Alternatively, one might imagine that the first two pieces of evidence are the opinion of a single expert or group of experts about the estimands y_1 and y_2. The researcher wants to recover the informative content in the expert's knowledge about τ contained in y_1, but is concerned that the manner in which they observe or interpret the causal process of interest is biased in some way that also biases their perception of ψ contained in y_2. The researcher can thus use an unbiased estimate of ψ from an experiment to estimate the experts' cognitive bias.

As before, the key question is whether the researcher can improve her inferences about the primary estimand of interest, τ, given her beliefs about the evidence-generating process. In all examples, the true $\tau = 1$, but I consider cases in which $\psi = 0$ and $\psi = -1$. Although both are constants in reality, the researcher has normal priors on both, allowing for correlation in the posterior beliefs about the estimands. The exclusion restriction is met in all pieces of evidence.

The table below describes the true values of the parameters under the two states of the world I consider ($\psi = 0$ and $\psi = -1$).

Note that because treatment effects are constant, there is no variance in the sampling error, since in all cases the SATE is equal to the PATE (see equation 22).

	$\psi = 0$			$\psi = -1$		
	y_1	y_2	y_3	y_1	y_2	y_3
θ_t	1.00	0.00	0.00	1.00	−1.00	−1.00
$\mathrm{E}[\epsilon^S]$	0.00	0.00	0.00	0.00	0.00	0.00
$\sqrt{\mathrm{Var}[\epsilon^S]}$	0.00	0.00	0.00	0.00	0.00	0.00
$\mathrm{E}[\epsilon^Z]$	0.11	0.11	0.00	0.11	0.11	0.00
$\sqrt{\mathrm{Var}[\epsilon^Z]}$	0.04	0.04	0.04	0.04	0.04	0.04
$\mathrm{E}[\epsilon^V]$	0.00	0.00	0.00	0.00	0.00	0.00
$\sqrt{\mathrm{Var}[\epsilon^V]}$	0.00	0.00	0.00	0.00	0.00	0.00
$\sqrt{\mathrm{Var}[\hat{\theta}_t]}$	0.04	0.04	0.04	0.04	0.04	0.04

Table 2: True values of the parameters in the simulation study. Features of the different kinds of evidence obtained by simulating implementation of the evidence-generating process in `DeclareDesign`. Evidence of type y_1 produces information on τ with some unknown bias β, evidence of type y_2 produces information on ψ with some unknown β, and evidence of type y_3 produces unbiased information on ψ.

5.1.1 Results

By repeating the three evidence-generating processes 200 times and computing the posterior distribution over the unknown parameters each time, we can estimate the expected improvement in the quality of the researcher's inferences about τ as she changes her evidence synthesis strategy. Table 3, figure 3 and figure 4 present the results. In the first row of the table and plots, the researcher considers only one piece of evidence of type y_1. In the second row, she considers two pieces of evidence: one each of type y_1 and y_2. In the third, she considers three pieces of evidence, one each of type y_1, y_2 and y_3. The columns of the table report the expected variance, the expected bias, and the expected mean squared error in the marginal posterior distribution over τ, for states of the world in which $\psi = 0$ and $\psi = -1$, respectively. The columns of the plots represent the densities of the *expected* marginal posterior distribution over τ, ψ and β^Z, with the true values plotted as solid vertical lines and the mean of the expected posterior distribution plotted as a dashed vertical line. In figure 3 $\psi = 0$, while in figure 4 $\psi = -1$.

	$\psi = 0$			$\psi = -1$		
Strategy	Post. Var.	Bias	MSE	Post. Var.	Bias	MSE
y_1 only	14.86	-0.45	15.23	14.90	-0.47	15.28
y_1 and y_2	10.33	-0.35	10.56	10.57	-0.08	10.69
y_1, y_2 and y_3	4.22	-0.15	4.28	4.23	-0.06	4.27

Table 3: Results of the three-study simulation experiment. All outcomes with respect to the marginal posterior distribution over τ.

Turning firstly to the meta-analysis that only uses one piece of evidence, the first thing to note in table 3 is that the expected variance in the marginal posterior distribution over τ is much greater than the

variance in estimates of τ generated by the frequentist estimator. This reflects prior uncertainty about the true value of the estimand and the unknown bias.

The expected mean of the posterior distribution is also biased towards 0, because the researcher is effectively averaging the expected likelihood estimate of approximately 1.1 and the prior mean of approximately 0 (see the dotted lines in the first column of figures 3 and 4 to observe this more clearly). The averaging of the estimate with the prior in these data-sparse cases coheres with the intuition outlined in section 2, in which new information on the bias term is required to improve inferences. With only one piece of evidence, y_1, the differences in inferences between the two states of the world are negligible, since information on ψ is never used to update on τ (this is why the density is missing from the middle column of the top row in the figures 3 and 4).

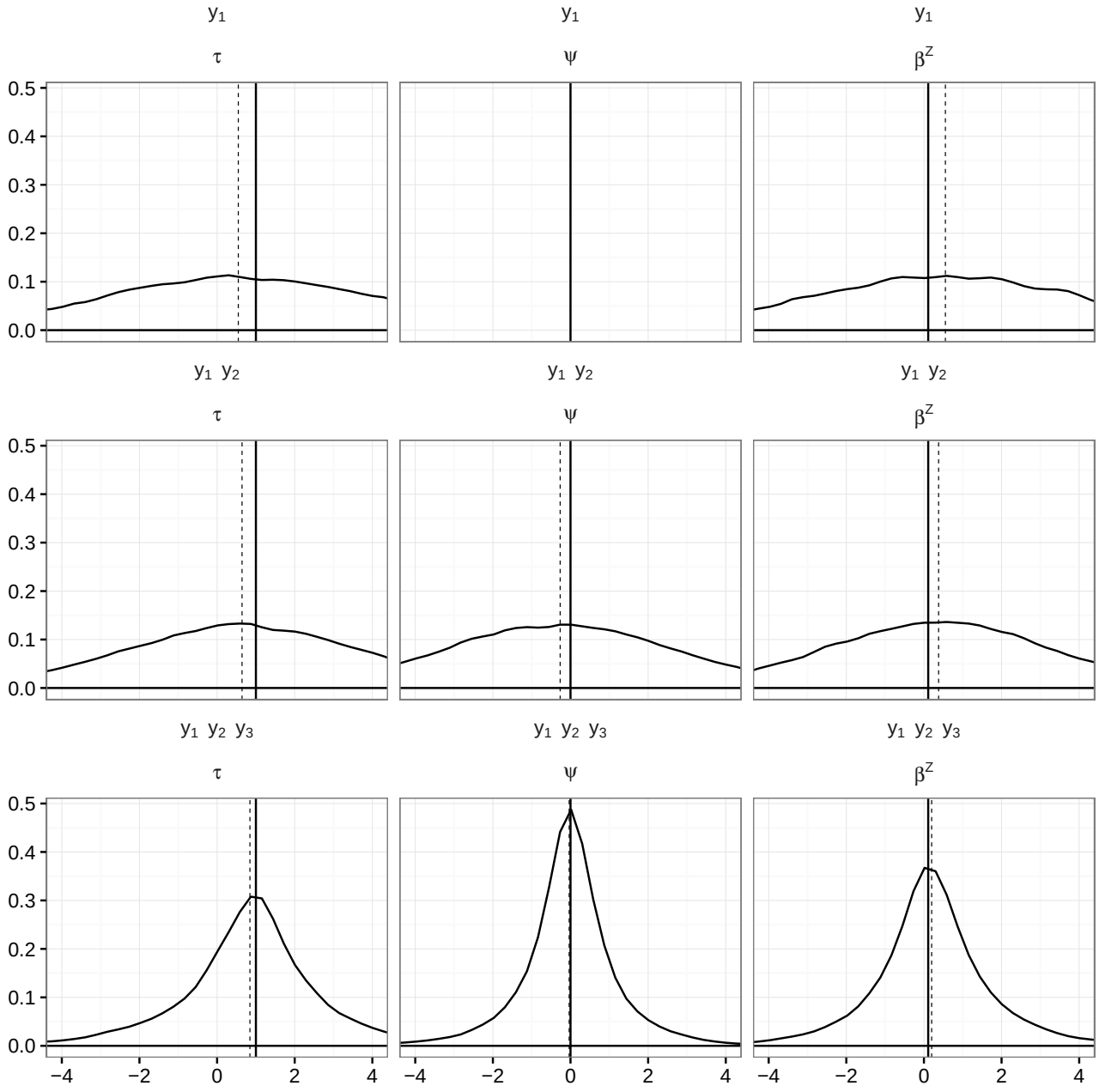

Figure 3: Marginal posterior densities over two estimands and one bias parameters, with $\psi = 0$. Vertical lines show true values, curved lines show marginal posterior density. Inferences in first row condition on one piece of evidence of type y_1, those in second row condition on two pieces of evidence, with one of type y_1 and y_2 each, and those in third row condition on three pieces of evidence, with one of type y_1, y_2 and y_3 each. Evidence of type y_1 produces information on τ with some unknown bias β, evidence of type y_2 produces information on ψ with some unknown bias β, and evidence of type y_3 produces unbiased information on ψ.

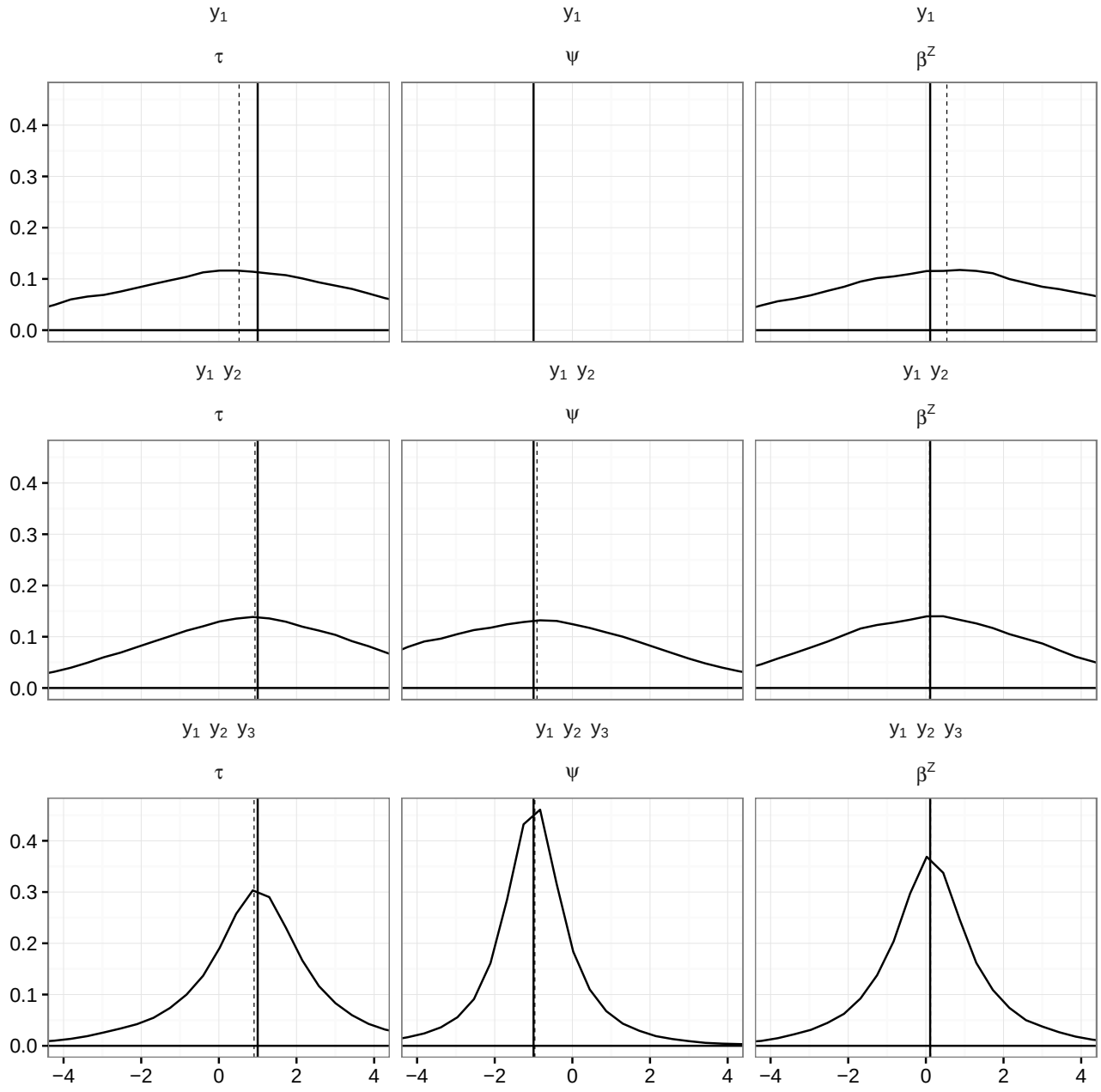

Figure 4: Marginal posterior densities over two estimands and one bias parameters, with $\psi = -1$. Vertical lines show true values, curved lines show marginal posterior density. Inferences in first row condition on one piece of evidence of type y_1, those in second row condition on two pieces of evidence, with one of type y_1 and y_2 each, and those in third row condition on three pieces of evidence, with one of type y_1, y_2 and y_3 each. Evidence of type y_1 produces information on τ with some unknown bias β, evidence of type y_2 produces information on ψ with some unknown bias β, and evidence of type y_3 produces unbiased information on ψ.

Moving to the second row of table 3, irrespective of the true value of ψ, adding information of type y_2 to her evidence synthesis strategy improves the researcher's expected posterior mean squared error over τ, simultaneously reducing the expected variance and the expected bias. In the next section I attempt to parse out whether this variance reduction derives solely from the inclusion of more data in the meta-analysis per se, or whether the specific data type matters in terms of the marginal inferential improvement.

While the variance reduction may be attributed to a sample size increase, however, it is clear that the reduction in the expected bias in the posterior distribution over τ is a function of the true value of ψ and thus sensitive to the type of data. While in both cases the the expected posterior variance is reduced by roughly 30% through the inclusion of y_2, in the case that $\psi = -1$, expected bias in the posterior is reduced by roughly 80%, versus the roughly 25% reduction in bias that occurs when $\psi = 0$.

This experiment also illustrates the benefits of the triangulation strategy. As figures 3 and 4, the variance in the expected posterior density is much lower when the bias term can be identified off the inclusion of unbiased data on ψ. Relative to including only y_1, the evidence on the estimand of interest (τ), the inclusion of two pieces of evidence (y_2 and y_3) about a different estimand greatly improves inferences about the primary evidence. Specifically, in both states of the world the reduction in expected posterior variance is equivalent to approximately 75%. Moreover, the inclusion of unbiased evidence greatly reduces the expected bias, especially in the case where $\psi = 0$. These findings show that the intuitions presented in section 2 generalize to more realistic statistical modeling scenarios. One difference, however, is in the bias term: whereas the inclusion of three pieces of evidence was sufficient to reduce bias to approximately 0 when the data-generating process was assumed to be a simple sum of binary variables, here the expected bias in the posterior remains distribution is not equal to 0. This might be due to bias in the BEAMA approach, or due to the fact that there is insufficient data to identify the unknown parameters, so that the zero-centered prior is still pushing the posterior toward zero. I show in the next section that data sparsity, and not a fundamental flaw in the BEAMA approach, is behind this non-zero bias.

5.2 Medium-Sized Meta-Analysis with One Unknown Bias

The preceding simulation study showed that it is possible to learn from evidence at risk of bias provided that triangulation of the bias is possible. However, it was not able to show which kinds of evidence best improve inferences, and which combinations of evidence lead to the best inferences. Moreover, the non-zero bias even in the triangulated case is concerning: does this arise because the approach is flawed or because the uncertainty about unknown parameters allows the zero-centered prior to pull the posterior

downwards toward 0?

To answer this question, in this simulation analysis I vary the different proportions of the data types, holding other features of the first simulation study constant, and setting $\psi = 0$. This set of examples can be thought of as similar to meta-analysis settings in which researchers have 5 - 50 studies at their disposal, or as similar to a situation in which a large group of possibly biased experts are queried about a causal parameter.

I examine how inferences change as a function of different combinations of evidence of types y_1, y_2 and y_3. Table 4 outlines the parameters in the simulation analysis.

$N(y_1)$	1	5	1	1	5	5	1	10	1	1	5	10	5	10	1	5	1	10	5	5	10	10	1	10	10	5	10
$N(y_2)$	1	1	5	1	5	1	5	1	10	1	5	5	10	1	10	1	5	5	10	5	10	1	10	10	5	10	10
$N(y_3)$	1	1	1	5	1	5	5	1	1	10	5	1	1	5	5	10	10	5	5	10	1	10	10	5	10	10	10
N	3	7	7	7	11	11	11	12	12	12	15	16	16	16	16	16	16	20	20	20	21	21	21	25	25	25	30

Table 4: Parameters for simulation analysis. Evidence of type y_1 produces information on τ with some unknown bias β, evidence of type y_2 produces information on ψ with some unknown β, and evidence of type y_3 produces unbiased evidence on ψ.

The simulation study varies both the relative proportion of the different evidence types as well as the overall size of the meta-analysis. It is therefore possible to examine not only how inferences improve as any type of data is added, but also to examine the impact of adding one type of data, marginalizing across all combinations of the other types of data.

5.2.1 Results

Figure 5 marginalizes over the different proportions of evidence types to examine how inferences vary simply as a function of the total amount of data included. The first panel illustrates the expected variance in the marginal posterior distribution on τ, the second panel illustrates the expected absolute bias, and the third illustrates the expected mean squared error.

The first noteworthy feature of this analysis is the very sharp improvement in inferences that arises from increasing the total amount of data from 3 to 7 pieces of evidence. In these cases, the amount of two types of evidence is held constant at 1, while the amount of the other type of evidence is increased to 5 data points (see table 4). The marginal reductions in variance quickly drop to 0 as there are fewer types of evidence for which only one piece of information is available, effectively reducing to 0 as at least five pieces of each kind of evidence are included.

Bias is also reduced to zero, albeit more slowly than variance. Indeed, it is not until more than one piece of information of each different kind of evidence is available that the expected bias in the posterior reduces to 0. This provides an answer to the question above: the BEAMA approach is biased toward

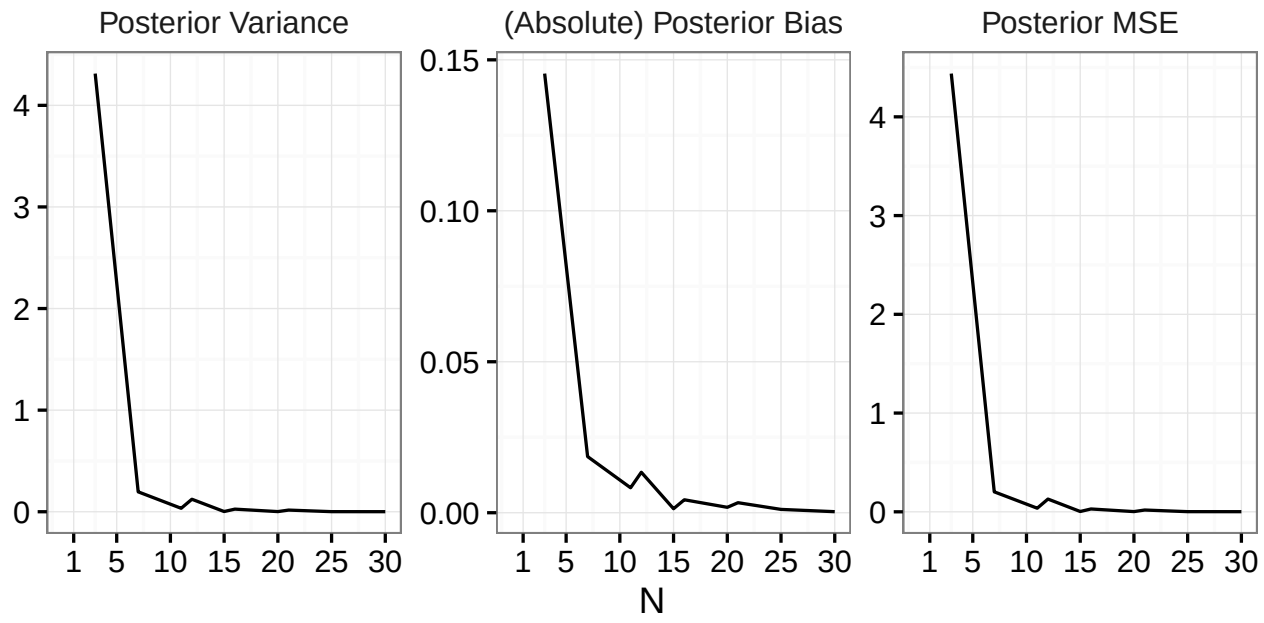

Figure 5: The marginal improvement in inferences about τ as the total number of studies in the meta-analysis increases.

the prior when only a very sparse amount of data is available. As soon as it is possible to identify the location of the three unknown parameters, however, triangulation is effective at reducing bias to 0.

Table 5 presents one of the most interesting and perhaps counter-intuitive aspects of BEAMA. The first three rows represent moving from a situation akin to the previous simulation analysis, in which only one piece of information of each evidence type is available, to one in which five pieces of evidence of one type become available. The last three rows represent moving from a situation in which the researcher holds five pieces of each kind of information, to one in which she holds ten pieces of one kind of information. The columns report the marginal reduction in expected posterior variance, absolute bias, and mean squared error that results from these shifts. Thus, the values tell us the marginal probative value of different kinds of information.

Intuitively, one might expect that the greatest marginal improvement in inferences on τ would come from data on τ. Alternatively, one might assume that, since unknown bias is the problem, collecting more unbiased evidence will best improve inferences. Indeed, this has shown to be the case for contexts in which it is possible to gather unbiased data on the estimand of interest (Gerber, Green, and Kaplan, 2004; Welton et al., 2009). However, in the case considered here, the greatest marginal improvement in inferences in fact comes from increasing the amount of data of type y_2, which is biased information on an estimand that is not primarily of interest.

The intuition behind this finding is as follows: in the case that the researcher has infinite data of type

	Reduction in posterior variance	Reduction in (absolute) bias	Reduction in MSE
With 1 each, increasing y_1 to 5	-4.145	-0.117	-4.263
With 1 each, increasing y_2 to 5	-4.173	-0.140	-4.294
With 1 each, increasing y_3 to 5	-4.031	-0.124	-4.149
From 5 each, increasing N(y_1) to 10	-0.001	-0.001	-0.001
From 5 each, increasing N(y_2) to 10	-0.001	-0.001	-0.001
From 5 each, increasing N(y_3) to 10	-0.001	-0.001	-0.001

Table 5: Marginal improvement in inferences from increasing data of one type at a given meta-analysis size. Each row shows improvement in inferences when increasing data points of evidence type y_i while each number of data points of other evidence types is held constant. Evidence of type y_1 produces information on τ with some unknown bias β, evidence of type y_2 produces information on ψ with some unknown β, and evidence of type y_3 produces unbiased information on ψ.

y_1 and y_3, she will never improve her inferences about τ because she is unable to infer anything informative about β from the combination of y_1- and y_3-type evidence. The y_2-type evidence provides information on terms common to both y_1- and y_3-type evidence, and is thus used more efficiently in the statistical likelihood. Table 5 also confirms the observation of decreasing marginal returns to data inclusion: when the researcher already has five pieces of each kind of information, her inferences are already good and do not improve greatly through the inclusion of five more pieces of any kind of evidence.

In table 5, including data of type y_2 reduces both the expected variance and the expected bias, suggesting that this kind of evidence dominates on both dimensions. However, figure 6 suggests that the superior marginal improvement derives mostly from how it affects the first moment of the expected posterior distribution over τ.

Each line illustrates the change in inferences as data of type i increases, averaging across all other possible amounts of other types of data. As the middle panel shows, whereas the marginal reduction in variance for a given type of data across all proportions of the other types of data is similar for y_1, y_2 and y_3, the marginal reduction in absolute bias is systematically greater as more evidence of type y_2 is included in the study, compared to including more y_1- or y_3-type evidence.

5.3 Medium-Sized Meta-Analysis with Incorrectly Specified Bias

The foregoing analyses paint an optimistic picture of the potential for BEAMA to improve inferences. However, they both are very optimistic in assuming that the first moments of the assignment error (the bias) are exactly the same in studies y_1 and y_2. This strong assumption is made by the model implemented in 4, but as we saw in section 2, it is not necessary for the BEAMA framework in general. A more flexible statistical model may simply state informative priors on the correlations among bias terms without stating

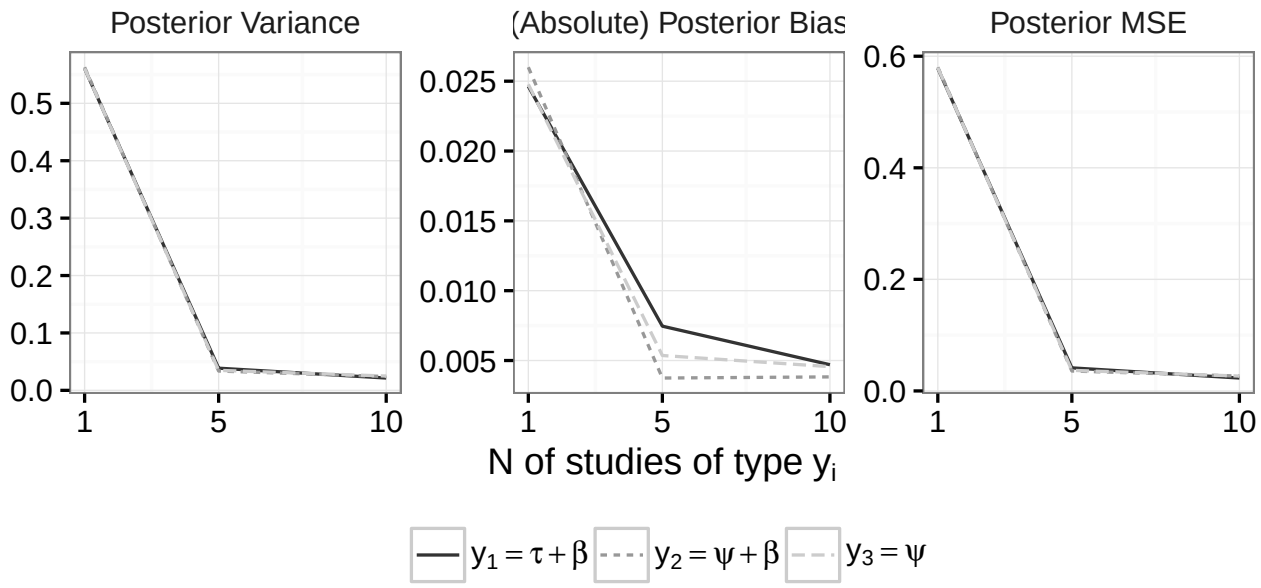

Figure 6: The marginal improvement in inferences as the number of pieces of evidence of type i is increased. Each line represents the expected variance, bias or mean-squared error in the posterior over τ given amount of evidence of type i, integrating over all other combinations of evidence of the other types. Evidence of type y_1 produces information on τ with some unknown bias β, evidence of type y_2 produces information on ψ with some unknown β, and evidence of type y_3 produces unbiased information on ψ.

they are identical. Nevertheless, it is worth investigating how wrong inferences can be if two different kinds of bias are incorrectly specified as being the same using the model developed in section 4.

In this simulation analysis, I use a very similar setup to the preceding studies. There are thirty pieces of information, 10 each of three kinds. While the researcher models y_1 and y_2 as sharing a common bias, however, in fact their biases are different. Table 6 presents the true values of the parameters. In the first three columns, the bias that affects y_1 is assumed to be smaller in magnitude to that which affects y_2, but is similarly positive. In the final three columns, the bias is of the same magnitude but is negative.

	$\overbrace{\beta_\psi^Z > \beta_\tau^Z}$			$\overbrace{\beta_\psi^Z < 0}$		
	y_1	y_2	y_3	y_1	y_2	y_3
θ_t	1	-1	-1	1	-1	-1
$\mathrm{E}[\epsilon^Z]$	.11	.22	0	.11	$-.11$	0
$\sqrt{\mathrm{Var}[\epsilon^Z]}$	.04	.04	.04	.04	.04	.04

Table 6: True values of the parameters in the simulation study. All other parameters held as in table 6. Note that the assignment bias is not the same across biased pieces of evidence.

5.3.1 Results

Recall that in a 30-study meta-analysis when the true bias was correctly specified as the same, the bias term reduced to 0. In this simulation, however, the expected bias remains non-zero, as illustrated in table 7.

Problem	Post. Var.	Bias	MSE
Larger positive bias	0	-0.11	0.01
Negative bias	0	0.22	0.05

Table 7: Results of the medium-sized meta-analysis simulation experiment with bias incorrectly specified as common. All outcomes with respect to the marginal posterior distribution over τ.

Perhaps unsurprisingly, the expected bias in the posterior is equivalent to the true difference in the biases. This arises because the BEAMA model assumes a functional form in which $\tau = y_1 - (y_2 - y_3)$, so that when the bias is not equal, the expected bias in the posterior is equal to $(\tau + \beta_\tau^Z - (\beta_\psi^Z + \psi - \psi)) - \tau = \beta_\tau^Z - \beta_\psi^Z$. As illustrated in section 2, it is in fact possible to improve inferences using BEAMA if the biases are simply correlated. In theory, if the parameter α in $\beta_\tau^Z - \alpha\beta_\psi^Z = 0$ could be estimated, then it would be possible to recover the unbiasedness of BEAMA, at the expense of increasing uncertainty through the introduction of α. I do not explore such methods in this paper, but plan to do so in future extensions.

5.4 Medium-Sized Meta-Analysis with Two Unknown Biases

The previous simulation analysis showed that mis-specifying bias as common across studies when in fact it is not can be inferentially damaging: researchers can become over-confident in the wrong answer. Moreover, real meta-analysis settings are likely to feature sources of evidence at risk from of bias from multiple sources. Thus, it will often be necessary to specify multiple unknown bias terms.

The last simulation study examines the potential for estimating multiple unknown biases by expanding the principle of triangulation into a network context. Consider the diagram in figure 7.

The arrows represent causal relationships. Here, for example, assignment bias β^Z and the primary causal effect of interest, τ, have a direct causal effect on the evidence of type y_1.

Here in principle we can no longer use y_1, y_2 and y_3 to triangulate τ independently from β^Z, because the third type of evidence is itself at risk of bias by β^V. This could arise, for example, if y_1 and y_2 represent estimates of τ and ψ_1 from an observational study, while y_3 is an estimate of ψ_1 from a second experimental study, however the researcher knows that participants were not blinded to treatment conditions in evidence of type y_3 and so the exclusion restriction may be violated. Suppose, however, that a third estimand ψ_2, was also estimated in the non-blinded RCT, and that a third study generates evidence on ψ_2 that is

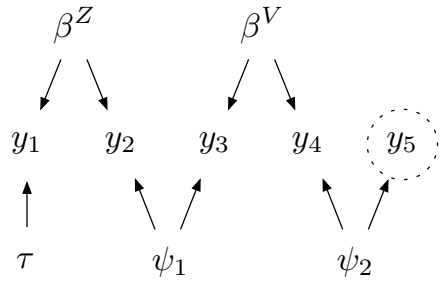

Figure 7: A networked meta-analysis allowing for triangulation of two sources of bias. The primary and two ancillary estimands are denoted τ, ψ_1 and ψ_2, respectively. Evidence of type y_1 and y_2 is at risk of assignment bias, while evidence of type y_3 and y_4 is at risk of exclusion restriction bias. Evidence of type y_5 is at risk of neither kind of bias. Evidence of type y_5 is encircled with a dotted line to indicate that its presence in the network varies across the two simulations in this analysis.

free of the assignment and the exclusion restriction biases in the other two studies. Then in principle y_5 allows triangulation of β^V providing information on ψ_1, which by extension enables triangulation of β^Z providing information on τ. The analysis seeks to understand the marginal impact of including the y_5 type of evidence on inferences about τ. Note that y_5 is twice-removed from τ: it shares no common parameters with any evidence on τ.

The simulation study investigates the marginal impact of including y_5 by generating simulated evidence similarly to the scenario outlined above, using the true parameters defined in table 8 below.

	$\overbrace{\quad}^{\tau}$	$\overbrace{\quad\quad}^{\psi_1}$		$\overbrace{\quad}^{\psi_2}$	
	y_1	y_2	y_3	y_4	y_5
θ_t	1	-1	-1	1	1
$\mathrm{E}[\epsilon^Z]$	.1	.1	0	0	0
$\mathrm{E}[\epsilon^V]$	0	0	$-.2$	$-.2$	0
$\mathrm{Var}[\epsilon^Z]$	.04	.04	.04	.04	.04
$\mathrm{Var}[\epsilon^V]$	.001	.001	.001	.001	.001

Table 8: True values of the parameters in two-bias simulation study. Assignment and exclusion bias vary across kinds of evidence.

I compare two cases. In the first, the researcher sets $N(y_1) = N(y_2) = N(y_3) = N(y_4) = 5$, and $N(y_5) = 0$. In the second, her evidence strategy is $N(y_1) = N(y_2) = N(y_3) = N(y_4) = N(y_5) = 4$. Thus, the total size of the meta-analysis is held constant at 20 pieces of evidence, but the proportions of evidence types included change.

5.4.1 Results

Table 9 and figure 8 present the results. Each pair of rows in the table illustrates the expected bias, variance and mean squared error in the posterior of the indicated estimand or bias parameter when the

y_5 data is or is not included in the meta-analysis. The figure illustrates the expected marginal posterior densities over the parameters, with the two strategies in the rows and the different parameters in the columns.

Parameter	Evidence strategy	Bias	Variance	MSE
τ	Without y_5	-0.008	2.179	2.847
τ	With y_5	0.001	0.004	0.007
ψ_1	Without y_5	0.008	2.174	2.823
ψ_1	With y_5	0.000	0.002	0.004
ψ_2	Without y_5	0.000	2.167	2.820
ψ_2	With y_5	0.000	0.001	0.001
β^Z	Without y_5	-0.048	2.783	3.584
β^Z	With y_5	-0.004	0.003	0.005
β^V	Without y_5	-0.107	5.588	7.244
β^V	With y_5	0.000	0.002	0.003

Table 9: Results of the medium sized meta-analysis with two unknown biases.

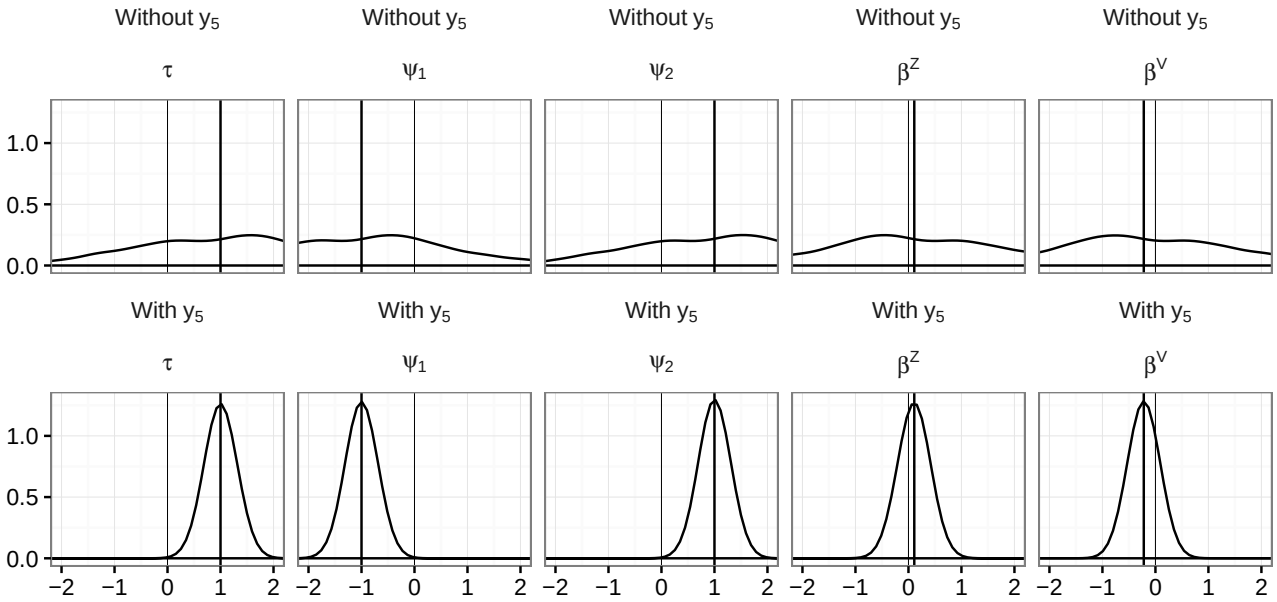

Figure 8: Marginal posterior densities over three estimands and two bias parameters, with and without evidence of type y_5. Vertical lines indicate the true value of the parameter, curved lines indicate marginal posterior densities. The first row features inferences only including evidence of type y_1, y_2, y_3, and y_4, while the second row includes these and evidence of type y_5.

As expected, the inclusion of the unbiased data strongly improves inferences, with the greatest marginal improvement on the unknown bias term. Interestingly, the marginal improvement on the other parameters takes place mainly through the reduction of variance. Even without any unbiased data, the expected posterior mean is very close to the true value of the estimands.

6 Discussion and Conclusion

I first discuss the main findings of the paper, before reviewing the limitations of the method it proposes and outlining ways I plan to address these in future work.

Key findings.

- Despite the finding in single-estimand contexts according to which including more data at risk of unknown bias does not improve inferences when researchers know nothing about the true scae and location of the bias, I find that including data at risk of unknown bias on a different estimand *does* improve inferences about the primary estimand, if and only if priors on estimands are not perfectly correlated, priors on bias terms are not perfectly independent, and there is some uncertainty in the posterior beliefs about the estimand.

- If two pieces of evidence contain information on different estimands, beliefs about bias terms are correlated, and the evidence is not perfectly informative about the estimand of interest, then the inclusion of an unbiased piece of evidence on the ancillary estimand improves inferences about the primary estimand when all three types of evidence are considered, *even if* the unbiased evidence does not itself contain any information about the primary estimand.

- When the bias term is correctly specified as common across two sources of evidence on different estimands, and a third piece of unbiased evidence on one of the estimands exists, increasing the amount of information reduces expected bias in inferences to 0, rendering the biased information perfectly informative.

- When different biases (or different estimands) are specified as common across two sources of evidence, BEAMA produces biased estimates of the estimand of interest.

- If the goal is to learn about a primary estimand but all available information on that estimand is biased, then the greatest expected marginal improvement in inferences is generated by including evidence that shares the bias term and an ancillary estimand about which unbiased information is available, relative to biased data on the primary estimand or unbiased data on the ancillary estimand.

- BEAMA is biased toward the prior in small meta-analyses, due to uncertainty around the first moment of the bias and estimand terms.

- If multiple sources of bias affect different sources of evidence on different estimands and if those bias terms are common to evidence on estimands about which at least one piece of unbiased evidence exists, then the unbiased evidence (possibly on some uninteresting estimand) is sufficient to reduce uncertainty and bias about the other estimands and bias terms.

Limitations. First and foremost among BEAMA's limitations is the bias correlation assumption, which is shown to be very damaging if violated in the way the model is setup in section 4. This assumption can be lent more empirical weight if it can be shown, for example, that estimates of some bias effect are homogeneous among a range of contexts. In this case researchers have empirical grounds for justifying the assumption that this bias will work similarly in the evidence whose informative content they wish to recover.

A second limitation in this paper is the assumption of additive bias on a common scale. Recall that bias is defined in section 3 as the expectation of an error arising from a difference in potential outcomes. Thus, by definition I ruled out multiplicative bias in the above. In reality, multiplicative forms of bias are conceivable, and outcomes are often expressed upon different scales. Ultimately the question of scaling is less damaging as various methods exist to standardize data or estimates to a common scale. The question of multiplicative bias, however, remains an open one. I point to it here as a possible scope condition on the claims made above.

Finally, I focused on sources of bias in the first moment of the difference-in-means estimator, but a wide literature on bias in variance estimates also exists. Conceivably, a researcher might want to take account of this in her meta-analysis. In principle, more direct modeling of bias in variance estimates could be incorporated.

Directions for Further Development. As a data collection and analysis strategy, BEAMA points to an interesting set of directions for applied evidence synthesis. Foremost among these is the notion that if the only information about an estimand of interest that exists is at risk of some bias, then it is possible to recover the informative content of that information by investing in identification of the bias term by leveraging ancillary estimands. In the medical and social sciences, there are frequently estimands about which it is difficult to gather unbiased evidence. For example, it can be ethically unfeasible to randomize certain life-saving treatments, or when it is feasible this can only be done on convenient samples from the population about which inferences are sought. Similarly, even when it is possible to develop unbiased evidence on some estimand, the sample size of such evidence might be very limited such that there is great uncertainty around the evidence. In such cases, if less noisy but possibly biased evidence exists,

and its bias can be triangulated through evidence on an ancillary estimand, then both the unbiased and biased evidence can be used together to reduce uncertainty.

Another potentially fruitful avenue for further development is the integration of qualitative expert information into quantitative causal inference. Gill and Walker (2005) called for social science researchers to make use of the rich information that experts hold in contexts in which data is sparse or external validity is low. They pointed to a range of potential applications for such expert data. Despite this, very few studies since have taken up this approach. One reason may be that expert data is at risk of unknown bias and therefore it is difficult to recover its informative content. However, in the "supra-bayesian" approach to elicitation (Winkler, 1968; Lindley, 1983; French, 1983; Roback and Givens, 2001), expert opinion is simply more evidence on an estimand that an analyst can integrate into a single statement of beliefs (a so-called "consensus prior"). Seen from this light, the distinction between expert elicitation and meta-analysis is trivial, and BEAMA could be used in conjunction with the consensus prior approaches developed recently, for example in Albert et al. (2012). If expert information is sought about a primary estimand, but unbiased information exists on an ancillary estimand about which they also have information, then their bias might be estimated by eliciting judgments on both the primary and ancillary estimands. Moreover, there may be instances in which cognitive bias can be estimated directly through the use of survey experiments, in which case it is treated as its own ancillary estimand.

As outlined above, benchmarking the uniformity of bias across evidence sources seems an important step in assessing the broader applicability of BEAMA. Future work refining the BEAMA statistical model may point to ways of grouping non-equal but similar bias into clusters in situations where bias is found to be heterogeneous across evidence sources.

* * *

This paper proposed bias estimation adjusted meta-analysis (BEAMA), a method for synthesizing possibly biased sources of evidence in order to make inferences about one or several estimands of interest. It formalized the method and provided intuitions through simple algebraic examples, developed these intuitions into a partially-known hierarchical gaussian probability model based on a decomposition of the additive error in a difference-in-means estimator, and tested this model's performance in simulated meta-analyses comprised of data generated through simulated sampling, assignment and estimation mechanisms. As a first step in a procedure that is still not well-defined, the method yet has several limitations, none of which cannot be addressed through further development. The method arguably points to new avenues for meta-analysts that wish to recover the informative content of possibly biased evidence without relying on the subjective opinion of experts.

References

Albert, Isabelle, Sophie Donnet, Chantal Guihenneuc-Jouyaux, Samantha Low-Choy, Kerrie Mengersen, and Judith Rousseau. 2012. "Combining Expert Opinions in Prior Elicitation." *Bayesian Analysis* 7 (3): 502–532.

Aronow, Peter M, Donald P Green, Donald KK Lee et al. 2014. "Sharp bounds on the variance in randomized experiments." *The Annals of Statistics* 42 (3): 850–871.

Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2016. "Stan: A probabilistic programming language." *Journal of Statistical Software* .

Chaimani, Anna, Haris S Vasiliadis, Nikolaos Pandis, Christopher H Schmid, Nicky J Welton, and Georgia Salanti. 2013. "Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study." *International journal of epidemiology* 42 (4): 1120–1131.

Darvishian, Maryam, Maarten J Bijlsma, Eelko Hak, and Edwin R van den Heuvel. 2014. "Effectiveness of seasonal influenza vaccine in community-dwelling elderly people: a meta-analysis of test-negative design case-control studies." *The Lancet Infectious Diseases* 14 (12): 1228–1239.

Dias, Sofia, NJ Welton, VCC Marinho, G Salanti, JPT Higgins, and AE Ades. 2010. "Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173 (3): 613–629.

Eddy, David M, Vic Hasselblad, and Ross Shachter. 1990. "A Bayesian method for synthesizing evidence: the confidence profile method." *International journal of technology assessment in health care* 6 (01): 31–55.

Eddy, David M, Victor Hasselblad, Ross D Shachter et al. 1992. *Meta-analysis by the confidence profile method*. Academic Press.

Efthimiou, Orestis, Thomas Debray, Gert Valkenhoef, Sven Trelle, Klea Panayidou, Karel GM Moons, Johannes B Reitsma, Aijing Shang, and Georgia Salanti. 2016. "GetReal in network meta-analysis: a review of the methodology." *Research synthesis methods* .

French, Simon. 1983. *Group consensus probability distributions: A critical survey*. University of Manchester. Department of Decision Theory.

GAO. 1992. *Cross design synthesis: a new strategy for medical effectiveness research.* US Government Accounability Office.

Gelman, Andrew et al. 2006. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis* 1 (3): 515–534.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. "The illusion of learning from observational research." In *Problems and Methods in the Study of Politics*, ed. Ian Shapiro, Rogers M Smith, and Tarek E Masoud. Vol. 28 Cambridge University Press pp. 251–273.

Gill, Jeff, and Lee D Walker. 2005. "Elicited priors for Bayesian model specifications in political science research." *Journal of Politics* 67 (3): 841–872.

Hoffman, Matthew D., and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15: 1351 – 1381.

Humphreys, Macartan, and Alan M Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109 (04): 653–673.

Imai, Kosuke, Gary King, and Elizabeth A Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the royal statistical society: series A (statistics in society)* 171 (2): 481–502.

Kaizar, Eloise E. 2015. "Incorporating both randomized and observational data into a single analysis." *Annual Review of Statistics and Its Application* 2: 49–72.

Kynn, Mary. 2008. "The 'heuristics and biases' bias in expert elicitation." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (1): 239–264.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating random correlation matrices based on vines and extended onion method." *Journal of multivariate analysis* 100 (9): 1989–2001.

Lindley, D. V. 1983. "Reconciliation of probability distributions." *Operations Research* 31: 866–880.

McCarron, C Elizabeth, Eleanor M Pullenayegum, Lehana Thabane, Ron Goeree, and Jean-Eric Tarride. 2011. "Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance." *PloS one* 6 (10): e25635.

Roback, Paul J, and Geof H Givens. 2001. "Supra-Bayesian pooling of priors linked by a deterministic simulation model." *Communications in Statistics-Simulation and Computation* 30 (3): 447–476.

Schmitz, Susanne, Roisin Adams, and Cathal Walsh. 2013. "Incorporating data from various trial designs into a mixed treatment comparison model." *Statistics in medicine* 32 (17): 2935–2949.

Spiegelhalter, David J, and Nicola G Best. 2003. "Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling." *Statistics in medicine* 22 (23): 3687–3709.

Splawa-Neyman, Jerzy, DM Dabrowska, TP Speed et al. 1990. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science* 5 (4): 465–472.

Thompson, Simon, Ulf Ekelund, Susan Jebb, Anna Karin Lindroos, Adrian Mander, Stephen Sharp, Rebecca Turner, and Désirée Wilks. 2011. "A proposed method of bias adjustment for meta-analyses of published observational studies." *International journal of epidemiology* 40 (3): 765–777.

Turner, Rebecca M, David J Spiegelhalter, Gordon Smith, and Simon G Thompson. 2009. "Bias modelling in evidence synthesis." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172 (1): 21–47.

Tversky, Amos, and Daniel Kahneman. 1975. "Judgment under uncertainty: Heuristics and biases." In *Utility, probability, and human decision making*. Springer pp. 141–162.

Welton, NJ, AE Ades, JB Carlin, DG Altman, and JAC Sterne. 2009. "Models for potentially biased evidence in meta-analysis using empirically based priors." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172 (1): 119–136.

Winkler, R. L. 1968. "The consensus of subjective probability distributions." *Management Science* 15: 361–375.

Wolpert, Robert L, and Kerrie L Mengersen. 2004. "Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke." *Statistical Science* pp. 450–471.

7 Appendix

A MCMC Convergence

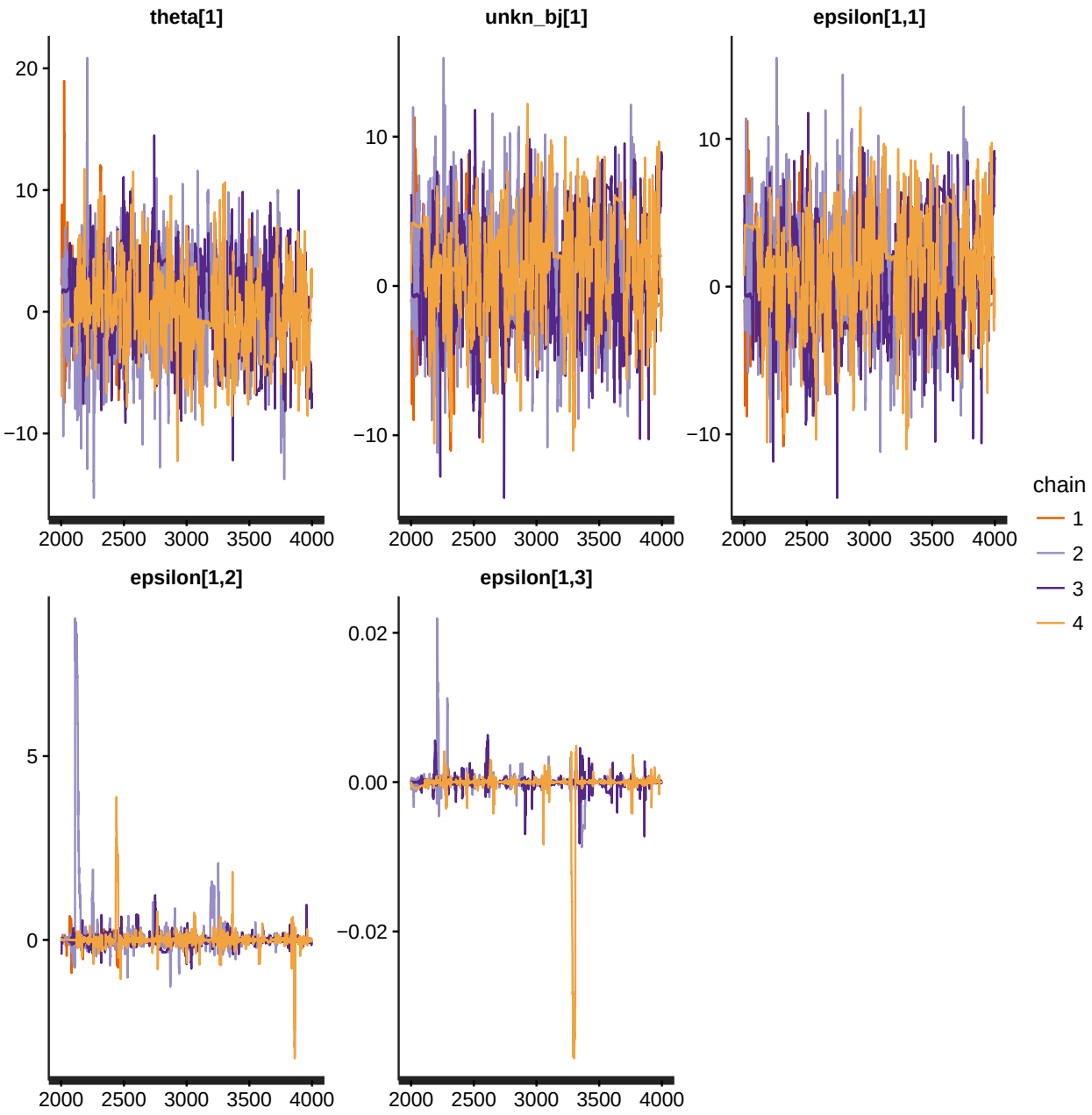

Figure 9: Traceplots of estimand, unknown bias and error parameters in the MCMC samples produced by Stan.

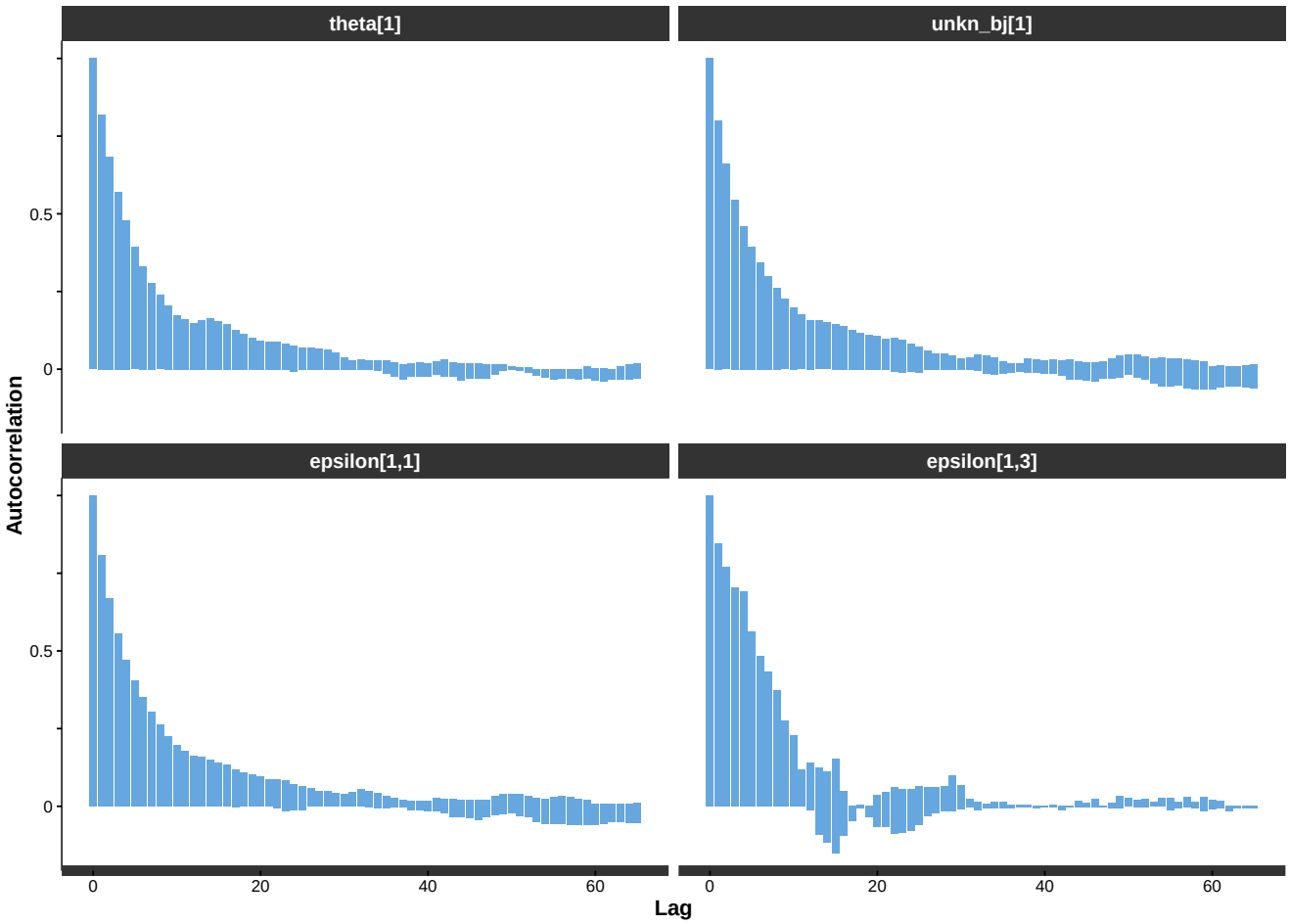

Figure 10: Autocorrelation in samples of estimand, unknown bias and error parameters in the MCMC draws produced by Stan.