PSYCHOMETRIKA—VOL. 82, NO. 2, 459–474
JUNE 2017
DOI: 10.1007/s11336-016-9507-z

CAUSAL INFERENCE FOR META-ANALYSIS AND MULTI-LEVEL DATA STRUCTURES, WITH APPLICATION TO RANDOMIZED STUDIES OF VIOXX

MICHAEL SOBEL AND DAVID MADIGAN

COLUMBIA UNIVERSITY

WEI WANG

PHILIPS RESEARCH NORTH AMERICA

We construct a framework for meta-analysis and other multi-level data structures that codifies the sources of heterogeneity between studies or settings in treatment effects and examines their implications for analyses. The key idea is to consider, for each of the treatments under investigation, the subject's potential outcome in each study or setting were he to receive that treatment. We consider four sources of heterogeneity: (1) response inconsistency, whereby a subject's response to a given treatment would vary across different studies or settings, (2) the grouping of nonequivalent treatments, where two or more treatments are grouped and treated as a single treatment under the incorrect assumption that a subject's responses to the different treatments would be identical, (3) nonignorable treatment assignment, and (4) response-related variability in the composition of subjects in different studies or settings. We then examine how these sources affect heterogeneity/homogeneity of conditional and unconditional treatment effects. To illustrate the utility of our approach, we re-analyze individual participant data from 29 randomized placebo-controlled studies on the cardiovascular risk of Vioxx, a Cox-2 selective nonsteroidal anti-inflammatory drug approved by the FDA in 1999 for the management of pain and withdrawn from the market in 2004.

Key words: causal inference, individual participant data, meta-analysis, multi-level models, randomized experiment, research synthesis.

1. Introduction

Meta-analyses are widely used in educational, behavioral, and medical science. In most applications, data from published papers are used. Here, the information available to account for systematic sources of heterogeneity in treatment effects across studies is generally inadequate, and researchers will often estimate a common treatment effect, and possibly the variation in this effect across studies (DerSimonian & Laird, 1986; Raudenbush, 2009). Individual participant data (IPD) meta-analyses, of primary focus herein, offer numerous advantages (Cooper & Patall, 2009; Simmonds et al., 2005), and are becoming increasingly common (as evidenced, for example, by the many papers listed on the website of the Cochrane Individual Participant Meta-Analysis Methods Group). The simplest analytic strategy is to pool the subject level data from all studies, treat the pooled data as if it comes from a single study, and estimate a common treatment effect, as for example, in Ross et al. (2009). But average treatment effects frequently vary across studies. Subjects in different studies are often drawn from different populations. Heterogeneity in average treatment effects may also stem from grouping treatments with different effects, as when doses z

Electronic supplementary material The online version of this article (doi:10.1007/s11336-016-9507-z) contains supplementary material, which is available to authorized users.

For helpful comments on a previous draft of this paper, we are grateful to Eva Petkova, Yajuan Si, three anonymous reviewers, and the editor. Sobel's research was supported by NIH grant R01EB016061.

Correspondence should be made to Michael Sobel, Department of Statistics, Columbia University, New York, NY, USA. Email: michael@stat.columbia.edu

and z^* ($z \neq z^*$) of a drug, administered respectively in studies s and s^* ($s \neq s^*$), are treated as a single treatment z. Administrative or contextual differences between studies, as when a classroom intervention is administered by different teachers or a therapy by different psychologists, or the long-term effects of a medical intervention are assessed in countries with different medical systems, may also lead to heterogeneous effects. Even if average treatment effects are the same in all studies, estimates from nonrandomized studies that do not adequately account for the treatment assignment mechanism may incorrectly suggest otherwise. The use of different outcome measures in different studies, incorrectly deemed commensurable after transformation, may also induce heterogeneity (Goldstein, Yang, Omar, Turner, & Thompson, 2000). Although we only explicitly consider the case where outcomes are measured on a common scale herein, our framework is applicable to outcomes that are commensurable after transformation.

To account for between-study heterogeneity in treatment effects, the introduction of covariates and/or random effects is often recommended. (Aitkin, 1999; Higgins, Whitehead, Turner, Omar, & Thompson, 2001; Higgins, Thompson, & Spiegelhalter; 2009; Tudur Smith, Williamson, & Marson, 2005). Unfortunately, researchers often use random effect models simply as a convenient device to obtain a parsimonious summary of otherwise disparate results, without concern for sources of between-study heterogeneity. In this case, as well as the case where fixed effects for studies and/or study by treatment interactions are estimated instead, we believe that more careful consideration of sources of heterogeneity can inform analyses and/or resulting policy recommendations. To that end, we develop a framework that explicitly codifies the sources and nature of between-study heterogeneity in meta-analyses. When extensive information about the subjects is collected, as in some individual participant data (IPD) meta-analyses, the framework can be used to test, under specified conditions, whether or not particular sources contribute to between-study heterogeneity.

In meta-analyses based on published data, subjects cannot be linked to covariates that vary within studies. If these characteristics differentiate outcomes and their distribution varies across studies, it will not be possible to reliably test hypotheses about these sources of variation in general (Petkova, Tarpey, Huang, & Deng, 2013). Nevertheless, the framework may still be used to think more carefully about the sources of heterogeneity and how one might want to conduct and interpret empirical analyses.

Although we develop the framework in the context of meta-analysis, it is important to note it will also be useful in many other contexts featuring hierarchical data structures, as when a randomized or conditionally randomized experiment is conducted across multiple sites or an educational intervention administered in different classrooms and/or schools.

We proceed as follows. Section 2 sets out a framework for causal inference in meta-analyses. We extend potential outcomes notation, used in the statistical literature on causal inference to represent the counterfactual nature of the causal relation (Imbens & Rubin, 2015), to apply to studies in addition to treatments, and we use the notation to define unit, conditional and unconditional average treatment and study effects. Hong and Raudenbush (2006), who study students nested within schools, use a similar notation. In their context, asking whether a student's outcome would be the same if he or she were enrolled in a different school is not of interest. Here, however, a subject's potential responses to the same treatment in different studies is key. We use this in codifying the sources of heterogeneous treatment effects, and further, when study effects are known to be 0, that is, when a subject's potential responses do not vary over studies, to test for differential selection into studies; the latter usage is similar in spirit to the utilization of known effects to test for hidden bias in Rosenbaum (1989). Conversely, if there is no differential selection into studies, we can test whether or not conditional study effects are 0, that is, whether or not subject's outcomes depend on a study. In Sect. 3, we illustrate the utility of the framework, using the Cox proportional hazards model (Cox, 1972) to reanalyze individual participant data from 29 randomized studies conducted by Merck & Co, Inc. to assess Vioxx, a COX-2 selective nonsteroidal anti-inflammatory drug approved by the FDA in 1999 for the treatment of osteoarthritis, acute pain in adults, and the relief of menstrual symptoms, and withdrawn from the market in 2004. Our analysis complements previous meta-analyses of these data, finding that Vioxx increases cardiac adverse events relative to placebo. We also include data from a high-dose treatment arm not included in prior studies, finding evidence for a dose–response relationship. Section 4 concludes.

2. Meta-analysis: A Causal Framework

To codify the sources of between-study heterogeneity in average treatment effects, we first define the set of potential outcomes each subject would have under any treatment in any study. These are used to introduce consistency conditions under which subject's responses are invariant over studies, develop ignorability conditions for the mechanisms assigning subjects to treatments and selecting subjects into studies, and define equivalent treatments. Implications of these conditions and definitions for analyses are then considered.

2.1. Notation and Estimands

Let A_s denote the set of treatments in study s, and $A \equiv \bigcup A_s$, with elements 1, ..., L, the collection of all treatments considered in the G studies. Let i = 1,...,n index the subjects in the G studies, Z_i, taking values $z \in A = (1, \ldots, L)$, the treatment to which unit i is exposed, S_i, taking values $s \in C = (1, ..., G)$, the study to which i is allocated, and $Y_i \equiv Y_i(S_i, Z_i)$ unit i's response on an outcome of interest. Let $\mathbf{X}_i$, with values $\mathbf{x} \in \Omega$, denote a vector of observed covariates. In general, $\mathbf{X}_i$ includes characteristics of subjects $\mathbf{V}_i$ that vary within studies and characteristics $\mathbf{W}(S_i)$ with the same value for all subjects in a given study, for example, a common indication in a medical study; in meta-analyses based on summary published data, the investigator does not observe the covariates $\mathbf{V}_i$.

In the literature on causal inference, it is standard to consider the outcomes a subject would have had under treatments other than that to which he/she was assigned. However, to codify and evaluate the sources of heterogeneous treatment effects in meta-analysis, it is critical to also consider the outcomes a subject would have had for a given treatment in studies other than the study in which he/she participated.

Let $\mathbf{z} = (z_1, ..., z_n)$, where $z_i \in A, i = 1, ..., n$, and $\mathbf{s} = (s_1, ..., s_n)$, where $s_i \in C, i = 1, ..., n$, and let $Y_i(\mathbf{s}, \mathbf{z})$ denote the response subject i would have had under the allocation $\mathbf{s}$ to studies and assignment $\mathbf{z}$ to treatments. Although we assume potential outcomes are defined for all treatment by study combinations, the notation and results are easily modified to handle the case where potential outcomes are not defined for all treatment by study combinations.

Above, the potential outcomes of subject i; hence, the effect of treatment z vs. z' for that subject, may depend on treatment assignments and allocations of other subjects, impacting the definition of unit effects, and the definition, identification, and estimation of average effects. In the causal inference literature, the stable unit treatment value assumption (SUTVA) (Rubin, 1980) is often made; under this assumption, treatments are well defined (do not have multiple versions) and a subject's outcomes depend only upon the treatment he/she is assigned, not the treatments to which other subjects have been assigned. We extend this assumption as follows:

A1 Extended stable unit treatment value assumption (ESUTVA): For all possible assignments $\mathbf{z}$ and allocations $\mathbf{s}$, $Y_i(\mathbf{s}, \mathbf{z}) = Y_i(s_i, z_i) \equiv Y_i(s, z)$.

Although neither SUTVA nor ESUTVA is testable, the assumption is often reasonable, as in the Vioxx example herein, and in many other contexts, for example, multicenter trials. However, when study participants interact, as in social networks, schools and neighborhoods, this assumption may require modification (Hong & Raudenbush, 2006; Sobel, 2006).

Next, we assume random sampling of the observed outcomes and covariates in each study and treatment group.

A2 Study sampling assumption: For all subjects i in study s, $s = 1, ..., G$, the random vectors $Y_i, \mathbf{X}_i \mid S_i = s, Z_i = z$ are independent and identically distributed as $Y, \mathbf{X} \mid S = s, Z = z$.

Assumption (A2) is used to model the distribution of the observed responses, conditional on treatment, study and covariates. ESUTVA and (A2) jointly imply $F(y \mid S = s, Z = z, \mathbf{X} = \mathbf{x}) = F(y(s, z) \mid S = s, Z = z, \mathbf{X} = \mathbf{x})$. Hereafter, ESUTVA and (A2) are assumed throughout.

The causal estimands of interest herein are functions H of the distributions $F(y(s, z) \mid S = s, \mathbf{X} = \mathbf{x})$, for example,

$$E(Y(s, z) - Y(s, z') \mid S = s, \mathbf{X} = \mathbf{x}), \tag{1}$$

the average effect of treatment z vs. z' in study s in the subpopulation $\mathbf{X} = \mathbf{x}$, or the relative effect (defined in terms of logged survival probabilities) at time t of treatment z vs. z' in study s in the subpopulation $\mathbf{X} = \mathbf{x}$,

$$\frac{\ln(1 - F_{Y(s,z)}(t \mid S = s, \mathbf{X} = \mathbf{x}))}{\ln(1 - F_{Y(s,z')}(t \mid S = s, \mathbf{X} = \mathbf{x}))}, \tag{2}$$

where $Y_i(s, z)$ $(Y_i(s, z'))$ is subject i's survival time in study s under treatment z (z'). Unconditional effects are also typically of interest, e.g., $E(Y(s, z) - Y(s, z') \mid S = s)$.

Study effects may also be defined, for example, the average effect of study s vs. s' at covariate value $\mathbf{x}$ and treatment z in the study population s'':

$$E(Y(s, z) - Y(s', z) \mid \mathbf{X} = \mathbf{x}, S = s''). \tag{3}$$

Also of interest are conditions under which effects are homogeneous (heterogeneous) across studies; an understanding of these can be used to inform hypotheses and conclusions about the sources of heterogeneity, if any. Between-study heterogeneity in treatment effects stems from a) differences in responses of a given unit to the same treatment in different studies, b) grouping different treatments with different effects, and c) the assignment mechanism(s) by which treatments and studies are paired with subjects. We examine these sources of variation and the implications of these for conducting meta-analyses.

2.2. *Response Consistency Assumptions*

We first formalize the notion that a subject's response to a given treatment is the same in all studies. This idea (or one of the less stringent versions below), which cannot be properly expressed without consideration of the outcomes subjects would have had in studies other than those in which they participated, is implicit in meta-analyses, where a common treatment effect (or conditional effect) is assumed to hold across studies. However, note that none of the consistency assumptions below implies that conditional average treatment effects are homogeneous across studies.

A3a Strong response consistency assumption for treatment z: For all studies s, s' and subjects i, $Y_i(s, z) = Y_i(s', z)$.

When assumption (A3a) holds for all treatments $z \in A$, we say the responses are strongly consistent. [Assumption (A3a) should not be confused with the assumption of consistency sometimes used in network meta-analyses, as in Higgins et al. (2012).] While assumption (A3a) may not hold for all treatments, it may hold for at least one treatment, for example, if subjects' responses to placebo are the same in all studies. For any treatment for which (A3a) holds, the study effect (3) is 0, and if (A3a) holds for all treatments, there are no study effects. Assumption (A3a) cannot be tested directly, as only one potential outcome per subject is observed. It is also overly strong, in particular stronger than required for identifying and estimating the effects previously considered. We therefore relax it as follows:

A3b Weak response consistency assumption for treatment z: For all s, s' and $\mathbf{X}$,

$$F(y(s, z) \mid S = s', \mathbf{X} = \mathbf{x}) = F(y(s', z) \mid S = s', \mathbf{X} = \mathbf{x}), \tag{4}$$

that is, within any study s', the conditional distribution of outcomes $Y(s, z)$ and $Y(s', z)$, $s \neq s'$, are identical.

When assumption (A3b) holds for all treatments $z \in A$, we say the responses are weakly consistent. This assumption is also not directly testable, but when subjects are randomized to treatments within studies (see the unconfoundedness assumption (A6)) and there is no differential selection into studies (see assumption (A7)), it is possible to test this assumption.

Assumption (A3b) can be further weakened to apply to treatment effects, as vs. responses:

A4 Weak consistency of effects of treatment z versus z': Given H, for all s, s' and $\mathbf{X}$, the causal estimands

$$\begin{aligned} &H(F(y(s, z) \mid S = s', \mathbf{X} = \mathbf{x}), F(y(s, z') \mid S = s', \mathbf{X} = \mathbf{x})) \\ &\quad = H(F(y(s', z) \mid S = s', \mathbf{X} = \mathbf{x}), F(y(s', z') \mid S = s', \mathbf{X} = \mathbf{x})). \end{aligned} \tag{5}$$

Note that assumption (A4) is H specific. For example, if responses in different studies are translated ($Y_i(s, z) = Y_i(1, z) + \alpha_s$) and (1) is the estimand of interest, (A4) will hold, but will not necessarily hold for functions $H' \neq H$. When (A4) holds for all pairs of treatments we say the treatment effects are weakly consistent. As above, assumption (A4) cannot be tested directly.

Although the consistency assumptions cannot be tested directly, substantive considerations can often be used to evaluate their plausibility. For example, Landoni et al. (2013) conducted a meta-analysis to compare the effect of several volatile anaesthetics with total IV anaesthesia on mortality after cardiac surgery. As the surgeries are performed by different doctors in different hospitals with different policies regarding length of stay, follow-up intervals, etc., one might expect a subject's survival time to differ across studies; in this case, one would expect (although it is not mathematically necessary) treatment effect variation across studies. On the other hand, in our meta-analysis of the randomized Vioxx studies, assumption (A3b) is quite reasonable; in this case (as we later discuss), if average treatment effects varied over studies, it would suggest that subjects in different studies are sampled from different populations.

2.3. Treatment Equivalence Assumptions

The assumption that two or more treatments can be grouped and treated as identical in their implications for the response is often made in applications. We now formalize this

A5a Strong equivalence of treatments z_1 and z_2 in study s: For all i, $Y_i(s, z_1) = Y_i(s, z_2)$.

This states that every subject's response to treatment z_1 is identical to his response to treatment z_2. As before, a weaker version suffices:

A5b Weak equivalence of treatments z_1 and z_2 in study s: $F(y(s, z_1) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y(s, z_2) \mid S = s, \mathbf{X} = \mathbf{x})$.

If treatments z_1 and z_2 are strongly (weakly) equivalent in all studies, treatments z_1 and z_2 are said to be strongly (weakly) equivalent.

If the strong response consistency assumption (A3a) and assumption (A5a) hold, $Y_i(s, z_1) = Y_i(s', z_2)$ for all i, s and s'; however, if assumption (A3a) or one of its weaker forms holds but assumption (A5a) (or one of its weaker forms) is made in error, differences between potential outcomes $Y_i(s, z_1)$ and $Y_i(s', z_2)$ will be incorrectly attributed to study heterogeneity. We briefly illustrate this in our analysis of the Vioxx studies. Heterogeneity of treatment effects may also be attributed incorrectly to treatment heterogeneity when the strong or weak equivalence assumption holds. For example, Covey (2007) performed a meta-analysis of randomized studies conducted to assess the effectiveness of therapies under different presentation formats (relative risk, absolute risk, and number needed to treat). After controlling for the fact that some studies used doctors as participants, others not, she found heterogeneity in treatment effects across studies, attributing these to minor variations in the manner in which formats were presented in the different studies. However, it is possible these variations were not responsible for the heterogeneity, and that the heterogeneity was due, for example, to differences in characteristics of study participants in different studies, not adequately accounted for; we discuss this more extensively later (see assumption (A7) about the selection of participants into studies).

2.4. Treatment Assignment and Selection into Studies

2.4.1. Treatment Assignment Mechanism Within each study, we assume that treatment assignment is independent of potential outcomes, given covariates $\mathbf{X}_1$:

A6 Unconfounded treatment assignment given observed covariates: for every s, and treatment $z \in A_s$, $F(y(s, z) \mid Z = z, S = s, \mathbf{X}_1 = \mathbf{x}_1) = F(y(s, z) \mid S = s, \mathbf{X}_1 = \mathbf{x}_1)$.

Assumptions (A6) and (A1) allow identification of the potential outcome distributions from the observed outcome distributions:

$$F(y \mid Z = z, S = s, \mathbf{X}_1 = \mathbf{x}_1) = F(y(s, z) \mid Z = z, S = s, \mathbf{X}_1 = \mathbf{x}_1). \tag{6}$$

If all studies are randomized, as in Thase et al. (2005), treatment assignment is unconfounded, both unconditionally and given covariates. In an observational study, treatment assignment will not generally be unconfounded, but may be unconfounded given covariates. However, assumption (A6) is not directly testable, and if made in error, estimates of treatment effects that use this assumption will be biased.

For a single two-arm study, a standard assumption used to nonparametrically identify treatment effects is that treatment assignment is strongly ignorable, given covariates (Rosenbaum & Rubin, 1983); this consists of the unconfoundedness assumption and the positivity assumption that on the support of $\mathbf{X}_1$, the probability of assignment to either arm is greater than 0. We do not make a positivity assumption here. Consider a study with a placebo and treatments A and B, with subjects classified as low or high risk and treatment A targeted at low risk subjects, and B at high risk subjects. To learn the effect of A (B) vs. placebo for low (high) risk subjects, one could randomly assign half the low risk subjects and half the high risk subjects to the control group, and the other low (high) risk subjects to treatment A (B). In this case, low (high) risk subjects have probability 0 of receiving treatment B (A) and the effect of A (B) on high (low) risk subjects is not nonparametrically identified; however, as these effects are not of interest, it would not make sense to assign high (low) risk subjects to A (B).

2.4.2. Study Selection Mechanism The assumption (A6) that treatment assignment is unconfounded given covariates addresses the allocation of subjects within studies. We now consider the allocation of subjects to studies. If the distribution of outcomes is unrelated to study selection, as would be the case if each study sampled the same population, $F(y(s, z) \mid S = s) = F(y(s, z) \mid S = s')$; in this case, we say that study selection is unconditionally unconfounded. However, different studies generally sample different populations [for example, Kivimaki et al. (2012)].

Nonetheless, observed covariates $\mathbf{X}_2$ may account for differential selection into studies, that is, studies and potential outcomes are independent, given $\mathbf{X}_2$:

A7 *Unconfounded study selection, given observed covariates: For all studies s, s' and treatments z, $F(y(s, z) \mid S = s, \mathbf{X}_2 = \mathbf{x}_2) = F(y(s, z) \mid S = s', \mathbf{X}_2 = \mathbf{x}_2)$.*

Like the consistency assumptions, the notion of unconfounded study selection cannot be properly formalized without considering the outcomes subjects would have had in studies other than those in which they actually participated. Note also that if subjects in each study are a sample from a population $\mathcal{P}$, study selection is unconditionally unconfounded and unconfounded given covariates $\mathbf{X}_2$; however, in general, unconfounded study selection given $\mathbf{X}_2$ does not imply study selection is unconditionally unconfounded, nor does unconditionally unconfounded selection imply unconfounded selection given $\mathbf{X}_2$.

Because the treatment assignment and study selection mechanisms may involve different covariates, we have allowed different sets of these, $\mathbf{X}_1$ and $\mathbf{X}_2$, to account for treatment assignment and study selection. However, both (A6) and (A7) will often hold with a common set of covariates. An important case occurs when each of the studies is completely randomized. Suppose assumption (A7) holds with covariates $\mathbf{X}_2 = (\mathbf{W}(S), \mathbf{V}_2)$, where $\mathbf{W}(S)$ and $\mathbf{V}_2$ vary between and within studies, respectively. Then, as (A6) also holds with covariates $\mathbf{X}_2$, treatment assignment and study selection are unconfounded, given $\mathbf{X}_2$. A special case occurs when each study is a random sample from a population $\mathcal{P}$; treatment assignment and study selection are then unconfounded unconditionally.

Another important case occurs when (A6) holds with covariates $\mathbf{X}_1$ and each study is a random sample from a population $\mathcal{P}$. As (A7) then holds with covariates $\mathbf{X}_1$, treatment assignment and study selection are unconfounded, given $\mathbf{X}_1$.

More generally, suppose (A7) holds with $\mathbf{X}_2$ above and (A6) holds with covariates $\mathbf{V}_1$, and note that (A6) will also hold with $\mathbf{X}_1 = (\mathbf{W}(S), \mathbf{V}_1)$. Then if $\mathbf{V}_1 = \mathbf{V}_2$, we have $\mathbf{X}_1 = \mathbf{X}_2$, so that (A6) and (A7) will hold with the same set of covariates. If the covariates in $\mathbf{V}_2$ are included in $\mathbf{V}_1$ and the additional covariates in $\mathbf{V}_1$ are related to the potential outcomes, but not treatment assignment, as might occur if an investigator were interested in variation in treatment effects across subpopulations, (A6) may nevertheless hold with $\mathbf{X}_1 = \mathbf{X}_2$; for example, this would be the case in a randomized experiment. It may frequently also be reasonable to pool the two sets $\mathbf{V}_1$ and $\mathbf{V}_2$. Although neither (A6) nor (A7) must hold in this case, the rationale is that variables in $\mathbf{V}_2$ but not $\mathbf{V}_1$ will be related to potential outcomes and selection into studies, but not treatment assignment (if these variables were related to treatment assignment, then, as they are related to potential outcomes, they should be included in $\mathbf{V}_1$); as such, it is hoped (A6) will hold with $\mathbf{X}_1 = (\mathbf{W}'(S), \mathbf{V}'_1, \mathbf{V}'_2)'$. Similarly, variables in $\mathbf{V}_1$ but not $\mathbf{V}_2$ will be related to potential outcomes and treatment assignment, but not study selection, so it seems reasonable to believe (A7) might hold with $\mathbf{X}_2 = \mathbf{X}_1 \equiv \mathbf{X}$.

2.5. Combining Assumptions: Empirical Implications

We now examine the implications of the foregoing assumptions for conducting meta-analyses. Throughout this section, we maintain the ESUTVA assumption (A1), the sampling assumption (A2), and the unconfounded treatment assignment assumption (A6) with $\mathbf{X}_1 = \mathbf{X}$. Then, as noted earlier, $F(y(s, z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y \mid Z = z, S = s, \mathbf{X} = \mathbf{x})$. We also assume $\mathbf{X}_2 = \mathbf{X}$ in assumption (A7).

Although neither the response consistency assumptions (A3a) and (A3b) nor the selection assumption (A7) are testable, if assumption (A3b) and (A7) (or the stronger set (A3a) and (A7)) hold, for every study s in which treatment z is administered, the distributions of the response, conditional on $\mathbf{X}$, are identical:

$$\begin{aligned}F(y \mid Z=z, S=s, \mathbf{X}=\mathbf{x}) &= F(y(s,z) \mid S=s, \mathbf{X}=\mathbf{x}) = F(y(s,z) \mid S=s', \mathbf{X}=\mathbf{x})\\ &= F(y(s',z) \mid S=s', \mathbf{X}=\mathbf{x}) = F(y \mid Z=z, S=s', \mathbf{X}=\mathbf{x}),\end{aligned} \tag{7}$$

where the first equality follows from (A6), the second from (A7), the third from (A3b) and the fourth from (A6).

Thus, as assumptions (A1), (A2) and (A6) are assumed to hold, if (7) fails to hold, at least one of assumptions (A3b) or (A7) must be incorrect; note that if assumption (A3b) is incorrect, clearly assumption (A3a) is incorrect. As previously noted, it will often be possible to assess the plausibility of the consistency assumptions by considering the way outcomes are measured in different studies, and when it is reasonable to maintain assumption (A3b), one therefore obtains a test of the selection assumption (A7). On the other hand, if an investigator is willing to maintain the selection assumption (A7), as when subjects in different studies are sampled from a common population, a test of the weak consistency assumption (A3b) is obtained using (7); if assumption (A7) is maintained and (7) fails to hold, any between-study heterogeneity in treatment effects, if present, stems from unit heterogeneity in the potential outcomes across studies.

Whereas assumptions (A4) and (A7) imply that the conditional treatment effects do not vary across studies, i.e., given H,

$$\begin{aligned}&H(F(y(s,z) \mid S=s, \mathbf{X}=\mathbf{x}), F(y(s,z') \mid S=s, \mathbf{X}=\mathbf{x})) =\\ &H(F(y(s',z) \mid S=s', \mathbf{X}=\mathbf{x}), F(y(s',z') \mid S=s', \mathbf{X}=\mathbf{x})),\end{aligned} \tag{8}$$

assumptions (A3b) and (A7) imply the stronger condition that the response distributions (7) are identical.

We now consider the weak equivalence assumption (A5b). For all studies in which treatments z_1 and z_2 are administered, this can be empirically assessed by testing equality of $F(y \mid S = s, Z = z_1, \mathbf{X} = \mathbf{x}) = F(y(s, z_1) \mid S = s, \mathbf{X} = \mathbf{x})$ and $F(y \mid S = s, Z = z_2, \mathbf{X} = \mathbf{x}) = F(y(s, z_2) \mid S = s, Z = z_2, \mathbf{X} = \mathbf{x})$. If the hypothesis of equality is accepted, this may suggest that it is reasonable to extend the assumption to studies where only one of the treatments is administered.

However, in meta-analyses when treatments z_1 and z_2 are not both administered in any study, as in two-arm studies comparing these treatments with placebo z_0, the weak equivalence assumption is not directly testable. However, if both the weak response consistency assumption (A3b) and assumption (A7) that study selection is unconfounded given covariates hold, weak equivalence can be assessed by testing the equality of $F(y \mid S = s, Z = z_1, \mathbf{X} = \mathbf{x}) = F(y(s, z_1) \mid S = s, \mathbf{X} = \mathbf{x})$ and $F(y \mid S = s', Z = z_2, \mathbf{X} = \mathbf{x}) = F(y(s', z_2) \mid S = s', \mathbf{X} = \mathbf{x})$.

3. Vioxx and Cardiovascular Risk: A Meta-analysis of the Merck Studies

Vioxx is a COX-2 selective, nonsteroidal anti-inflammatory drug (NSAID) that was approved by the FDA in May 1999 for the relief of signs and symptoms of osteoarthritis, the management of acute pain in adults, and the treatment of menstrual symptoms. Compared to standard NSAIDs like naproxen and ibuprofen, the COX-2 class of drugs offered the promise of pain relief with reduced risk of gastrointestinal side effects. However, studies later showed that Vioxx caused an array of cardiovascular thrombotic side effects such as myocardial infarction, stroke, and unstable angina, leading to its withdrawal from the market in 2004.

Several other meta-analyses assessing the cardiovascular risk of Vioxx have been conducted. Using 18 randomized studies, Jüni et al. (2004) conducted a study level meta-analysis with myocardial infarction (MI) as outcome, comparing subjects on Vioxx (grouping doses of 12.5 milligrams per day (mg/d), 25 mg/d, and 50 mg/d) with subjects on placebo, naproxen, and other NSAID's, finding Vioxx significantly increased the risk of MI. No study heterogeneity was found in the relative risk, nor did subgroup analyses suggest a dose–response relationship or heterogeneity in the effects of Vioxx compared with different types of controls. Kearney et al. (2006) conducted a study level meta-analysis, using 121 randomized trials to compare the effects of selective COX 2 inhibitors (Vioxx, etoricoxib, celecoxib, lumiracoxib, and valdecoxib), with placebo, finding COX-2 inhibitors significantly increased the risk of serious vascular events (MI, stroke or vascular death). No heterogeneity in the effects of the different COX-2 inhibitors was found, nor did the effect vary by whether or not the study permitted subjects to use aspirin or not. As the vast majority of studies used 25 mg/d doses of the Cox-2 inhibitors, it was not possible to evaluate the dependence of the response on dosage. In a study-level meta-analysis of 114 randomized trials, Zhang, Ding, and Song (2006) found that Vioxx increased the risk of renal events.

Meta-analyses with individual participant data have also been conducted. Early studies, which used Cox proportional hazards models, stratified on indications or studies and combined different doses of Vioxx, treating these as equivalent (Konstam et al., 2001; Reicin, Shapiro, Sperling, Barr, & Yu, 2002; Weir, Sperling, Reicin, & Gertz, 2003), concluded that Vioxx did not increase cardiovascular risk compared with placebo or other NSAIDs. Although an increased risk of Vioxx relative to naproxen was found, this was attributed to a possible cardioprotective effect of the latter. However, a subsequent analysis (Ross et al., 2009), using similar methods on a more comprehensive collection of the Merck studies, suggested an increased risk.

Table 1 lists the 29 studies, all of which were completed before the withdrawal of Vioxx from the market, included in our analysis. The outcome of interest is the time to an adverse cardiovascular event; Ross et al. (2009) provide further information on the events included and the rationale for their inclusion. Study indications are listed in column 2 and study duration is in the last column of the table; as in Ross et al. (2009), all studies are at least 4 weeks long. Although some of these studies incorporated arms for NSAIDs other than Vioxx, only the placebo and Vioxx arms are included in our analysis. Dosages included in the different Vioxx arms are displayed in column 3. The data from treatment arms with 5 mg/d (studies 033 and 068) and 175 mg/d (study 017) are not used in the analysis, as very few subjects were treated with these doses; in addition, the data from these arms essentially provide no information about the possible effects of treatment with these doses, as no adverse events were observed in these studies in either the control group or at these dosage levels.

In total, there are 14,641 subjects and 6676 subject-years at risk. However, the outcome is sparse (see columns 5 and 6); there are eight studies with no events in any arm and nine studies with only one adverse event. The sparsity makes it difficult to analyze the data study by study. Thus, meta-analysis is critical for synthesizing the totality of evidence.

3.1. Models and Results

The estimand of interest is the relative effect at time t of treatment z vs. z' in study s in the subpopulation $\mathbf{X} = \mathbf{x}$ given in Eq. (2).

In each study, subjects were followed for the duration of the study, unless an event occurred prior to the study end, in which case the event time was recorded; for subjects who did not experience an adverse event on study, it is only known that the survival time exceeds the study duration. Data of this form are said to be "type 1" censored. This is a form of independent (sometimes called noninformative) censoring in which the censoring time does not depend on the subject's survival time; relative to the case of informative censoring, where the censoring time is not independent of the survival time, analysis is simplified as the censoring mechanism does not need to be modeled.

TABLE 1.
Randomized placebo-controlled trials of 4 weeks or longer conducted by Merck and finished before the withdrawal of Vioxx from the market.

Trial number	Indication studied	Dosage (mg/days)	Duration (weeks)	Event counts for Vioxx (person-year)	Event counts for control (person-year)
010	Osteoarthritis	25/125	6	2 (16)	0 (7)
029	Osteoarthritis	12.5/25/50	6	3 (46)	0 (16)
033	Osteoarthritis	5/12.5/25	6	1 (66)	1 (9)
040	Osteoarthritis	12.5/25	6	2 (72)	0 (11)
044	Osteoarthritis	25/50	24	3 (154)	0 (52)
045	Osteoarthritis	25/50	24	3 (157)	3 (61)
058	Osteoarthritis	12.5/25	6	1 (21)	0 (6)
083	Osteoarthritis	25	64	0 (21)	1 (21)
085	Osteoarthritis	12.5	6	1 (61)	0 (28)
090	Osteoarthritis	12.5	6	5 (56)	0 (27)
112	Osteoarthritis	12.5/25	6	0 (104)	0 (15)
116	Osteoarthritis	25	6	1 (54)	0 (15)
136	Osteoarthritis	25	12	1 (95)	1 (201)
219	Osteoarthritis	12.5	6	0 (18)	1 (8)
220	Osteoarthritis	12.5	6	0 (18)	0 (8)
017	Rheumatoid Arthritis	125/175	6	1 (8)	0 (7)
068	Rheumatoid Arthritis	5/25/50	8	1 (49)	0 (24)
096	Rheumatoid Arthritis	12.5/25	12	4 (97)	0 (58)
097	Rheumatoid Arthritis	25/50	12	0 (137)	0 (62)
098	Rheumatoid Arthritis	50	12	0 (11)	1 (12)
103	Rheumatoid Arthritis	50	12	0 (44)	0 (45)
078	Alzheimer's Disease	25	208	73 (1628)	56 (1758)
091	Alzheimer's Disease	25	52	13 (369)	14 (381)
126	Alzheimer's Disease	25	52	11 (193)	7 (197)
118	Chronic Nonbacterial Prostatitis	25/50	6	0 (15)	0 (8)
120	Low Back Pain	25/50	4	1 (28)	0 (14)
121	Low Back Pain	25/50	4	0 (23)	0 (11)
125	Migraine Prophylaxis	25	12	0 (23)	0 (22)
129	Familial Adenomatous Polyposis	25	24	0 (3)	0 (4)

We assume Y has density f_Y. To estimate (2), we model the hazard function (or instantaneous failure rate)

$$\begin{aligned} h(t, s, z \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) &\equiv \lim_{\Delta t \to 0^+} \frac{\Pr(t \leq Y < t + \Delta t \mid Y \geq t, Z = z, S = x, \mathbf{X} = \mathbf{x})}{\Delta t} \\ &= \frac{f_Y(t \mid Z = z, S = s, \mathbf{X} = \mathbf{x})}{1 - F_Y(t \mid Z = z, S = s, \mathbf{X} = \mathbf{x})} \end{aligned} \tag{9}$$

using the (semi-parametric) Cox proportional hazards model (Cox, 1972) in which the hazard is specified as the product of an unspecified (semi-parametric) baseline hazard $h_0(\cdot)$ with a parametric component $\exp(\lambda^\mathsf{T}\mathbf{X})$, where $\lambda \in \Lambda$ is a vector of parameters to be estimated; under the random sampling assumption (A2) and the assumption that for any two elements λ_1 and λ_2 of Λ, $\lambda_1^\mathsf{T}\mathbf{x} = \lambda_2^\mathsf{T}\mathbf{x}$ for all $\mathbf{x} \in \Omega$ implies $\lambda_1 = \lambda_2$, the parameter vector λ is identified.

Under the proportional hazards model, the hazard ratio is

$$\frac{h(t, s, z \mid Z = z, S = s, \mathbf{X} = \mathbf{x})}{h(t, s, z' \mid Z = z', S = s, \mathbf{X} = \mathbf{x})} = \frac{\ln(1 - F_Y(t \mid Z = z, S = s, \mathbf{X} = \mathbf{x}))}{\ln(1 - F_Y(t \mid Z = z', S = s, \mathbf{X} = \mathbf{x}))}. \tag{10}$$

Under the additional assumptions (A1) and (A6),

$$F_Y(t \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = F_{Y(s,z)}(t \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = F_{Y(s,z)}(t \mid S = s, \mathbf{X} = \mathbf{x}); \tag{11}$$

thus, the effect (2) is identified and equal to the hazard ratio (10).

For details on the estimation of the model and more generally on survival analysis, see the excellent treatment in Kalbfleisch and Prentice (2002); for discussion of semi-parametric survival models, for the modeling of response times in psychological testing, see Bloxom (1985). All models were estimated using the `survival` package (Therneau, 2013) in `R` (version 3.0.0).

Although meta-analysts typically do not think explicitly about response consistency (or inconsistency), we have seen (Sect. 2.5) that confronting this issue head on can have important implications for the analysis. Here, we assume potential responses are weakly consistent. This assumption (A3b) is justified because a given dose of Vioxx should have the same effect on a subject irrespective of the study in which he is enrolled, and the CVT adverse events are coded in different studies by medical professionals using the same rules. Under this assumption, recall that any treatment by study interaction, if present, must be due to differential selection of subjects into studies. In general we prefer to think of this source of heterogeneity as a fixed effect, and thus we do not explicitly consider frailty models (Duchateau & Janssen, 2007) herein (although the same conclusions are obtained when the η_{sz} are treated as Gaussian random variables, results available upon request).

Table 2 lists the models fitted to the Vioxx data. For each model, we tested the proportional hazards assumption using the standard Schoenfeld residuals test (Schoenfeld, 1982); in no case did we reject this assumption at the 0.05 level. Thus, we do not report Cox models stratified by indication, as in several previous analyses (Konstam et al., 2001; Reicin et al., 2002; Ross et al., 2009); these investigators, however, did not include covariates in their analysis.

In model M1, the most general model considered herein, the log hazard for subjects in study s assigned to treatment Z with covariates $\mathbf{X}$ consisting of the study level variable "indication" and individual level variables V is specified as:

$$\log h(t, s, z \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = \log h_0(t) + \alpha_s + \mathbf{v}^T \boldsymbol{\beta} + \mathbf{z}^T \boldsymbol{\theta} + \mathbf{z}^T \boldsymbol{\Theta} \mathbf{v} + \eta_{sz}, \tag{12}$$

TABLE 2.
Models fit to the data.

	Model specifications for log hazards
M1	Treatment + study + covariates + treatment×covariates + treatment×study
M2	Treatment + study + covariates + treatment×covariates
M3	Treatment + study + covariates
M4	Treatment + covariates + indication
M5	Treatment + covariates
M5*i*	Treatment + indication
M6	Treatment

Covariates include age and gender. Treatment is an unordered categorical variable.

where the α_s parameters account for study heterogeneity in the baseline hazard and $\boldsymbol{\beta}$ is a parameter vector associated with the individual level covariates $\mathbf{V}$; note that the study level variable "indications" cannot be included in models incorporating the study parameters because the collection of study indicator vectors and indications vectors are collinear. The individual level covariates are age and sex. Since the distributions of these variables vary across studies, and because age is a strong predictor of CV risk and event rates differ by sex, especially at younger ages, it would not be reasonable to think the selection assumption (A7) might hold were these variables not included in the analysis. Initially race was also included, but later dropped due to lack of significance in model M1 or any of the models parametrically nested under M1 that we considered. It is important to note that model (12) implicitly extrapolates the hazard to $(s, \mathbf{v})$ off the support of $(S, \mathbf{V})$, suggesting the need to justify this and/or examine the distribution of these covariates in different studies for common support; here there is one small study (study 118) in which only males are enrolled, and one small study (study 58) in which subjects are 80 or older, and when the analysis is conducted with these studies deleted, the results are virtually identical. $\mathbf{Z}$ is a vector of indicator variables indexing the five dosage levels: $Z_1 = 1$ for receipt of a 12.5 mg/d dose, 0 otherwise, $\ldots$, $Z_4 = 1$ for receipt of a 125 mg/d, 0 otherwise; thus, the parameter vector $\boldsymbol{\theta}$ compares each treatment with placebo. The treatment effects may also depend on covariates and studies: $\boldsymbol{\Theta}$ is the parameter matrix associated with treatment-by-covariate interaction, and the η_{sz} parameters, defined on the set $T = \{(s, z) \mid \text{treatment } z \text{ is administered in study } s\}$, which are the treatment by study interactions.

Under the assumption (A3b) that responses are weakly consistent, the treatment by study interactions η_{sz} and the study parameters α_s are 0 if assumption (A7) holds, i.e., if study selection is ignorable, given the covariates. Assumptions (A4) and (A7) also imply $\eta_{sz} = 0$.

If assumption (A3b) holds, assumption (A4) holds for any H. Therefore, as the hazard ratio does not depend on the α_s parameters, we can test for ignorable study selection by comparing the fit of model M2, in which the conditional treatment effects do not depend on study (all η_{sz} are 0), to that of model M1; the likelihood ratio test does not suggest M1 fits the data better ($-2\log\lambda = 38.4$, df = 40, $p = .54$, where λ is the likelihood ratio); thus we do not reject the selection assumption (A7).

Model M3 is the special case of M2 with no treatment by covariate interaction; under M3, the treatment effects do not depend on sex and age. Nor does the likelihood ratio test of Model M2 versus M3 suggest that model M2 fits the data better ($-2\log\lambda = 9.31$, df = 8, $p = .32$).

Model M4 simplifies M3, setting the study parameters α_s to 0, and replacing the study indicators with a vector of indicator variables for the seven study indications (Osteoarthritis, Rheumatoid Arthritis, Alzheimer's, Adenomatous Polyposis, Lower Back Pain, Prostatitis, Migraine). The substitution of indications for study does not result in a significant loss of information ($-2\log\lambda = 27.8$, df = 22, $p = .18$); thus M4 is preferred to M3.

Model M4 is consistent with the weak consistency assumption (A3b) and the assumption (A7) that the covariates $\mathbf{X}$ account for differential selection into studies, as these jointly imply the hazard function (and hence the distribution) of the outcomes does not depend on the study.

Models M5 and M5*i* further simplify model M4. In M5, which is preferred to M4 ($-2\log\lambda = 0.81$, df = 6, $p = .99$), age and sex are included, but not indications. However, model M5*i*, which includes the indications variable, but not age and sex, fits substantially worse than M4 ($-2\log\lambda = 58.41$, df = 2, $p < .001$). Thus, to account for study selection, it is necessary to include sex and age, but not indications. Further, given the magnitude of the test statistic comparing M4 with M5*i*, it is apparent that model M6, in which study selection is unconditionally unconfounded, will also fit much worse than M4.

Estimates of the treatment effects and 95% confidence intervals under model M5 are: (1) 2.63 and (1.30, 5.30) for 12.5 mg/days; (2) 1.33 and (0.99, 1.78) for 25 mg/days; (3) 2.38 and (1.03, 5.50) for 50 mg/days; and (4) 14.0 and (3.18, 61.63) for 125 mg/days. While the estimated hazard

TABLE 3.
Total number of subjects and days-at-risk under placebo and treatment arms (mg/d).

	Placebo	Vioxx 12.5	Vioxx 25	Vioxx 50	Vioxx 125
Subjects	5451	2462	5181	1432	115
Days-at-risk	1,127,971	122,532	1,057,899	123,931	4402

ratios all exceed 1, pointing to Vioxx causing an increase in the risk of an adverse cardiovascular event, it is difficult to tell whether or not the effect is the same for the various dosage levels, as assumed by prior researchers who grouped the doses in their studies, treating different doses as equivalent. We address this now.

3.2. Dose Response

In the majority of previous work (e.g.,Ross et al., 2009), investigators analyzed the data only from the 12.5, 25, and 50 mg/days arms (as more than 90 % of the patient-weeks of observation in their data were contributed by subjects receiving the 25 mg/days dose), treating these doses as equivalent. Table 3 displays the number of subjects and days-at-risk in the placebo and treatment arms for the data we analyzed.

To examine the dose–response relationship, we first fitted a restricted version of model M5, named M5g, in which 12.5, 25, and 50 mg/days doses are grouped into a low treatment level, with 125 mg/days as the high level. Model M5g is consistent with the treatment equivalence assumption (A5b), when the doses less than or equal to 50 mg/d are equivalent. While model M5g is not rejected by comparison with model M5 at the .05 level, ($-2 \log \lambda = 4.58$, df = 2, $p = .10$), a dose–response relationship is suggested. Furthermore, the more restricted model M5g^* that groups 12.5, 25, 50, and 125 mg/days into a single treatment level is rejected in comparison with model M5g ($-2 \log \lambda = 4.88$, df = 1, $p = .03$).

To further explore the relationship between dose and response, we also fit two more restricted versions of model M5, M5ℓ in which dosage is linearly related to the log hazard, with coefficient 0.015 and standard error 0.004, and M5q, which also includes squared dose. Model M5ℓ is marginally preferred to M5 ($-2 \log \lambda = 6.16$, df = 3, $p = .10$), and model M5q is not preferred to M5ℓ ($-2 \log \lambda = 0.91$, df = 1, $p = .34$). The Akaike information criterion (AIC) (Akaike, 1974) can be used to choose between models M5g and M5ℓ, weakly favoring M5ℓ (a difference of 0.66 in AIC).

Finally, recall the discussion in the previous section, which showed that when nonequivalent doses are combined, a researcher might be led to conclude that treatment effects are heterogeneous across studies when in fact they are homogeneous, as here. That occurs with these data. We treated all doses other than placebo as a single treatment and performed the same sequence of model comparisons as above, leading to the selection of model M3, in which study parameters α_s are needed to account for heterogeneity in the hazard, that is, the covariates alone cannot account for heterogeneity. The results (not shown here) are available upon request.

4. Discussion

In this paper, the potential outcomes notation developed in the statistical literature on causal inference is used to construct a framework for meta-analysis that helps to clarify and empirically examine the sources of between-study heterogeneity in treatment effects. The key idea is to consider, for each of the treatments under investigation, the subject's potential outcome in each study were he to receive that treatment.

Our re-analysis of the Vioxx studies is guided by this framework. The studies are randomized, so treatment assignment is unconfounded given covariates (assumption (A6)). We also assume the responses are weakly consistent (assumption (A3b)). While the weak consistency assumption is not empirically testable by itself as only one of the potential responses is observed per subject, it is easy to think about this assumption and substantive considerations can often provide a compelling rationale for believing (or not believing) it. Covariates are introduced into the analysis to investigate possible differences in subpopulation treatment effects and to account for differential selection into studies (assumption (A7)). The null hypothesis of no study by treatment interaction is not rejected at the 0.05 level, implying the conditional treatment effects are homogeneous across studies; in addition, the null hypothesis that the study parameters are 0 is not rejected at the 0.05 level, implying the response distributions for each treatment do not vary by study, conditional on covariates.

When the response consistency assumption (A3b) holds and treatments have not been grouped (or the equivalence assumption (A5b) holds), any variation in study level treatment effects is due to the differential distribution of covariates across studies. In this case, if individual participant data are available and measured covariates account for the distribution of subjects across studies; hence, for treatment effects that are heterogeneous over studies, as in our analysis of the Vioxx data, a researcher may wish to describe the heterogeneity in study level effects across covariate distributions or average the conditional effects over a target distribution of the covariates that reflects a population of intended recipients. One caveat is in order, however: if there are unobserved variables associated with both the potential responses and covariate distributions that differ in the subpopulations of study participants and the target population, use of the study results to extrapolate to the target population may be misleading; this is the problem of external validity (Campbell, Stanley, & Gage, 1963).

When treatment effects are homogeneous (either conditionally or unconditionally), the effect of treatment a vs. b can be obtained by combining the effect of treatment a vs. c with the effect of treatment c vs. b. For example, in any studies s, s' and s'' in which treatment pairs (a, b), (b, c), and (a, c) are administered, respectively, if treatment effects, defined as in (1), are unconditionally homogeneous, $E(Y(s, a) - Y(s, b) \mid S = s) = E(Y(s', a) - Y(s', c) \mid S = s') + E(Y(s'', c) - Y(s'', b) \mid S = s'')$. When treatment effects are heterogeneous, this is no longer the case. More generally, letting $\tau(a, b)$ denote the average of the a vs b treatment effects over the set of studies $C(a, b)$ in which both treatments a and b are administered, with $\tau(a, c)$, $C(a, c)$, $\tau(b, c)$ and $C(b, c)$ defined analogously, unless $C(a, b) = C(a, c) = C(b, c)$, in general $\tau(a, b) \neq \tau(a, c) + \tau(c, b)$ when heterogeneity is present. This phenomenon, termed incoherence (Lumley, 2002) or inconsistency (Higgins et al., 2012), has motivated researchers to develop models for detecting and accounting for this situation. Although heterogeneity does not imply incoherence, incoherence nevertheless stems from the sources of heterogeneity identified here, as homogeneity implies coherence. As Higgins et al. (2012) point out, when incoherence is detected in a network meta-analysis, it is not entirely clear how the analysis should subsequently proceed. Although that is not of concern herein, as we have accounted for heterogeneity in the unconditional effects using covariates, we believe our framework might be used to give some guidance on how to handle incoherence more generally when this phenomenon is encountered, and in future research we intend to examine this issue further.

Although the construction of our framework is guided by the case of meta-analysis, the framework will also be applicable to other types of multi-level data structures, as when an experiment is administered at different sites, students are observed in different classrooms and schools, and respondents live in different neighborhoods. This would also appear to be a fruitful area for further work.

References

Aitkin, M. (1999). Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, *18*(17–18), 2343–2351.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika*, *50*(4), 383–397.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*(2), 165.

Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making*, *27*(5), 638–654.

Cox, D. R. (1972). Regression models and life-tables regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, *34*(2), 187–220.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Duchateau, L., & Janssen, P. (2007). *The frailty model*. New York: Springer.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society: Series C*, *49*(3), 399–412.

Higgins, J., Jackson, D., Barrett, J., Lu, G., Ades, A., & White, I. (2012). Consistency and Inconsistency in network meta-analysis: Concepts and models for multi-arm studies. *Research Synthesis Methods*, *3*(2), 98–110.

Higgins, J., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A*, *172*(1), 137–159.

Higgins, J., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, *20*(15), 2219–2241.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, *101*(475), 901–910.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in statistics, social, and biomedical sciences*. New York: Cambridge University Press.

Jüni, P., Nartey, L., Reichenbach, S., Sterchi, R., Dieppe, P. A., & Egger, M. (2004). Risk of cardiovascular events and rofecoxib: Cumulative meta-analysis. *The Lancet*, *364*(9450), 2021–2029.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley.

Kearney, P. M., Baigent, C., Godwin, J., Halls, H., Emberson, J. R., & Patrono, C. (2006). Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the Risk of Atherothrombosis? *Meta-analysis of Randomised Trials British Medical Journal*, *332*(7553), 1302–1308.

Kivimäki, M., Nyberg, S. T., Batty, G. D., Fransson, E. I., Heikkilä, K., Alfredsson, L., . . ., IPD-Work Consortium. (2012). Job strain as a risk factor for coronary heart disease: A collaborative meta-analysis of individual participant data. *The Lancet*, *380*(9852), 1491–1497.

Konstam, M. A., Weir, M. R., Reicin, A., Shapiro, D., Sperling, R. S., Barr, E., et al. (2001). Cardiovascular thrombotic events in controlled, clinical trials of rofecoxib. *Circulation*, *104*(19), 2280–2288.

Landoni, G., Greco, T., Biondi-Zoccai, G., Neto, C. N., Febres, D., Pintaudi, M., et al. (2013). Anaesthetic drugs and survival: A Bayesian network meta-analysis of randomized trials in cardiac surgery. *British Journal of Anaesthesia*, *111*(6), 886–896.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, *21*(16), 2313–2324.

Petkova, E., Tarpey, T., Huang, L., & Deng, L. (2013). Interpreting meta-regression: Application to recent controversies in antidepressants efficacy. *Statistics in Medicine*, *32*(17), 2875–2892.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–316). New York: Russell Sage Foundation.

Reicin, A. S., Shapiro, D., Sperling, R. S., Barr, E., & Yu, Q. (2002). Comparison of cardiovascular thrombotic events in patients with osteoarthritis treated with rofecoxib versus nonselective nonsteroidal anti-inflammatory drugs (ibuprofen, diclofenac, and nabumetone). *The American Journal of Cardiology*, *89*(2), 204–209.

Rosenbaum, P. R. (1989). The role of known effects in observational studies. *Biometrics*, *45*, 557–569.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Ross, J. S., Madigan, D., Hill, K. P., Egilman, D. S., Wang, Y., & Krumholz, H. M. (2009). Pooled analysis of rofecoxib placebo-controlled clinical trial data: Lessons for postmarket pharmaceutical safety surveillance. *Archives of Internal Medicine*, *169*(21), 1976–1985.

Rubin, D. B. (1980). Randomization analysis of experimental data–The Fisher randomization test–Comment. *Journal of the American Statistical Association*, *75*(371), 591–593.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.

Simmonds, M. C., Higgins, J. P., Stewart, L. A., Tierney, J. F., Clarke, M. J., & Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials*, *2*(3), 209–217.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, *101*(476), 1398–1407.

Thase, M. E., Haight, B. R., Richard, N., Rockett, C. B., Mitton, M., Modell, J. G., et al. (2005). Remission rates following antidepressant therapy with bupropion or selective serotonin reuptake inhibitors: a meta-analysis of original data from

7 randomized controlled trials. *The Journal of Clinical Psychiatry*, *66*(8), 974–981.

Therneau, T. M. (2013). A Package for Survival Analysis in S. Retrieved from http://CRAN.R-project.org/package=survival (package version 2.37-4).

Tudur Smith, C., Williamson, P. R., & Marson, A. G. (2005). Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine*, *24*(9), 1307–1319.

Weir, M. R., Sperling, R. S., Reicin, A., & Gertz, B. J. (2003). Selective COX-2 inhibition and cardiovascular effects: A review of the rofecoxib development program. *The American Heart Journal*, *146*(4), 591–604.

Zhang, J., Ding, E. L., & Song, Y. (2006). Adverse effects of cyclooxygenase 2 Inhibitors on renal and arrhythmia events. *Journal of the American Medical Association*, *296*(13), 1619–1632.

Manuscript Received: 23 MAY 2015
Published Online Date: 7 JUL 2016