Meta-Analysis: A Causal Framework, with Application to Randomized Studies of Vioxx.

Michael E. Sobel, David B. Madigan, Wei Wang *

*Michael E. Sobel (E-mail:michael@stat.columbia.edu) and David B. Madigan (E-mail:madigan@stat.columbia.edu) are professors and Wei Wang (E-mail:wwang@stat.columbia.edu) is a doctoral candidate, Department of Statistics, Columbia University, New York, NY 10027.

ABSTRACT

We 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. We consider four sources of heterogeneity: 1) response inconsistency, whereby a subject's response to a given treatment varies across different studies, 2) the grouping of non-equivalent treatments, where subjects' responses to two or more different treatments vary, but the responses are assumed to be the same in the different treatments and the treatments are combined, 3) non-ignorable treatment assignment, and 4) response related variability in the composition of subjects in different studies. We then examine the implications of these assumptions for heterogeneity/homogeneity of conditional and unconditional treatment effects. To illustrate the utility of our approach, we re-analyze data from 29 randomized placebo controlled studies of Vioxx on the cardio-vascular risk of Vioxx, a Cox-2 selective non-steroidal anti-inflammatory drug approved by the FDA in 1999 for the management of pain and withdrawn from the market in 2004.

Keywords: Causal Inference, Individual Patient Data, Meta-Analysis, Randomized Experiment, Research Synthesis, Vioxx

1 INTRODUCTION

Meta-analyses combine data from multiple studies $s = 1, ..., G$ to estimate treatment effects with greater precision. In the simplest case, a researcher may use estimates from published studies to estimate a common treatment effect. In studies using individual patient data, a researcher might pool the subject level data from all studies and treat the pooled data as if it comes from a single study (for example Ross et al. 2009). However, average treatment effects often vary across studies. Subjects in different studies are often drawn from different populations. Heterogeneity in average treatment effects may also stem from grouping distinct treatments, as when doses z 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 the long-term effects of a medical intervention are assessed in countries with different medical systems, may also lead to heterogeneous outcomes. Further, even if average treatment effects are the same in all studies, estimates from non-randomized studies that do not adequately account for the treatment assignment mechanism may incorrectly suggest the effects vary over studies. The use of different outcome measures in different studies, deemed commensurable through standardization, may also induce heterogeneity across studies; this raises additional issues beyond the scope of this paper.

To deal with between study heterogeneity in treatment effects, each study may be treated as a distinct entity (no meta-analysis). Alternatively, covariates might be used to try to account for heterogeneity. The use of random effects models, in which effects from different studies are treated as a draw from a common distribution, is also often recommended (DerSimonian and Laird 1986; Aitkin 1999; Higgins et al. 2001, 2009). Although the latter two approaches may also be combined, researchers often ignore covariate information and use random effects models in a purely pragmatic way to generate a one number summary of

otherwise disparate effects. This occurs especially in meta-analyses based on results reported in published studies, where subjects in different studies may be quite different, even demonstrably so, but the investigator may not have enough information to account for between study variation in treatment effects or simply does not use the available information (study features) to do so. Greenland (1994) criticizes this practice, arguing that important sources of heterogeneity, which should be included in models to account for between study heterogeneity, are thereby obscured.

However, when treatment effects from different studies are exchangeable (conditionally exchangeable), as would be the case if the effects or conditional effects were sampled from a common distribution, the mean, and more generally the distribution of the random effects might then be of substantive interest as indicative of effects that would occur if an intervention were implemented more broadly in the future. But this need not be the case. As an example, suppose that in each study students in classrooms randomized to receive/not receive an intervention are compared. Suppose outcomes depend on the teacher administering the treatment. If the treatment can be scaled up and administered to students similar to those in the study by similar teachers, future effects will be drawn from the distribution of effects in the meta-analysis. However, if future treatments will be applied by teachers dissimilar to those in the meta-analysis, the distribution of the random effects may be of considerably less interest, and prediction intervals for effects in future classrooms based on the random effects distribution are likely to be misleading.

While the example above suggests the importance of distinguishing between different sources of between study heterogeneity and considering the implications of these differences for modeling and inference, in the vast majority of papers and applications where between study heterogeneity is considered, these sources are treated as undifferentiated. Using potential outcomes, we propose a framework for the identification of treatment effects in meta-analyses that codifies the sources and nature of between study heterogeneity, using

the assumptions that are made about these sources and the populations to which inferences are desired, to develop appropriate models for combining data from different studies and to understand what assumptions are made about heterogeneity (homogeneity), often implicitly, by researchers conducting meta-analyses. For previous work using potential outcomes in the context of meta-analysis, see Li et al. (2011), who used principal stratification to study the relationship between assignment to chemotherapy, a surrogate marker and five year survival status in a meta-analysis of colon cancer trials.

We proceed as follows. Section 2 sets out a framework for causal inference in meta-analyses. We extend the potential outcomes notation, standard in the statistical literature on causal inference, to apply to studies as well as treatments, using this notation to define both unit treatment and study effects. The unit treatment effects may then be averaged within covariate levels and studies to define conditional treatment effects, or averaged within studies to define study level treatment effects. Sources of between study heterogeneity in treatment effects are codified; conditions under which these are homogeneous are also given. In section 3, we illustrate the utility of the framework, reanalyzing individual patient data from 29 randomized studies conducted by Merck & Company, Inc. to assess Vioxx, a COX-2 selective non-steroidal 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 by Merck 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 formalize 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 formalize consistency conditions under which subject's responses are invariant over studies, to develop ignorability conditions for the mechanisms assigning subjects to treatments and selecting subjects into studies, and to define equivalent treatments. Implications of these conditions and definitions are then considered.

2.1 Notation and Estimands

Let Y denote a response of interest, A_s the set of treatments considered in study s, and $A \equiv \bigcup A_s$, with elements $1, ..., L$, the collection of all treatments observed in the G studies. Let $i = 1,...,n$ index the subjects in the collection of studies, Z_i, taking values $z \in (1, \ldots, L)$, the treatment to which unit i is exposed, and S_i, taking values $s \in (1, ..., G)$, the study to which i is allocated. Let $\mathbf{X}_i$, with values $\mathbf{x}$ and range Ω, denote a vector of subject characteristics. In general, $\mathbf{X}_i$ includes characteristics of subjects $\mathbf{V}_i$ that vary within studies, as well as characteristics taking on the same value for all subjects in a given study, as for example, when all subjects in a clinical study have a particular indication; in meta-analyses based on summary published data, only the latter type of characteristic would be available to investigators.

Although a subject is observed under only one treatment in one study, we consider the potential outcomes subjects would have in all other studies and treatments. Let $\mathbf{z} = (z_1, ..., z_n)$, $\mathbf{s} = (s_1, ..., s_n)$, and let $Y_i(\mathbf{s}, \mathbf{z})$ denote the response subject i would have under the allocation $\mathbf{s}$ to studies and $\mathbf{z}$ to treatments. We extend the stable unit treatment value assumption of Rubin (1980):

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

Under assumption (A1), made throughout, a subject's set of potential responses does not depend on the assignments of other subjects to treatments or studies. While this assumption is often reasonable in clinical trials and medical studies, it may be less reasonable when study participants interact, as in social networks or neighborhoods (Hudgens and Halloran 2008; Sobel 2006).

Second, we assume:

A2. Study sampling assumption: For all subjects i in study s, $s = 1, ..., G$, the random variables $Y_i \mid S_i = s, Z_i = z, \mathbf{X}_i = \mathbf{x}$ are independent and identically distributed with distribution function $F(y \mid S = s, Z = z, \mathbf{X} = \mathbf{x})$.

The causal estimands of interest in this paper can be expressed as 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 sub-population $\mathbf{X} = \mathbf{x}$, or the relative effect (defined in terms of logged survival probabilities) at time t of treatment z vs. z' in study s in the sub-population $\mathbf{X} = \mathbf{x}$

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

where $Y_i(s, z)$ ($Y_i(s, z')$) is the survival time of subject i in study s under treatment z (z'). 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'':

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

Unconditional versions of these effects are also typically of interest; these are the special case above omitting $\mathbf{X}$, e.g., $E(Y(s,z) - Y(s,z') \mid S = s)$.

2.2 Response Consistency Assumptions

Between study heterogeneity in treatment effects stems from a) between study differences in unit responses to treatments, b) grouping 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. To begin, we formalize the idea that unit responses to a given treatment are the same in all studies:

A3a. Strong response consistency assumption for treatment z: For all 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. (The response consistency assumption 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 the subject's responses to placebo are the same in all studies. For any treatment for which assumption (A3a) holds, the study effect (3) is 0, and if assumption (A3a) holds for all treatments, there are no study effects. But even if assumption (A3a) holds for every treatment, this does not imply the treatment effects are homogeneous across studies, as subjects may be assigned non-randomly to treatments and/or differentially selected

into studies.

Assumption (A3a) cannot be tested directly, as only one potential outcome per subject is observed. The assumption is also stronger than needed for identifying and estimating effects such as those previously considered. We therefore relax it as follows:

A3b. Weak response consistency assumption for treatment z: For all s, s' and 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, in the population from which study s'' is drawn, the conditional distributions of potential outcomes $Y(s,z)$ and $Y(s',z)$ are identical.

When assumption (A3b) holds for all treatments, we say the responses are weakly consistent. While this assumption is also not directly testable, substantive considerations can often be used to evaluate the plausibility of assumptions (A3a) and (A3b). For example, Covey (2007) performs a meta-analysis on studies that assess the perceived effectiveness of therapies under different presentation formats (relative risk, absolute risk, number needed to treat). If a given format were presented in exactly the same manner in each study, it would also be reasonable to believe that a subject's response to that format would be identical in all studies. However, Covey (2007) reports minor variations in presentations across studies, rendering assumptions (A3a) and (A3b) less plausible. Under additional conditions (discussed subsequently), it is possible to test assumption (A3b).

Assumption (A4) weakens assumption (A3b), applying the weak consistency assumption to the treatment effects, as versus responses:

A4. Weak consistency of effects of treatment z versus z': For all s, s' and s'' and $\mathbf{X}$,

the treatment effects

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

When assumption (A4) holds for all pairs of treatments we say the treatment effects are weakly consistent. As above, assumption (A4) cannot be tested directly. Note also that weak consistency of treatment effects does not imply between study homogeneity in treatment effects.

2.3 Treatment Equivalence Assumptions

Next, we consider the grouping of treatments:

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

As before, a weaker version suffices:

A5b. Weak equivalence of treatments z_1 and z_2 in study s: For all 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) holds and assumption (A5a) is made in error, differences between potential outcomes $Y_i(s,z_1)$ and $Y_i(s',z_2)$ will be incorrectly at-

tributed to study heterogeneity. Similar remarks apply with respect to the assumptions of weak consistency and weak equivalence. Under additional conditions, discussed later, we can assess whether or not grouping treatments that do not satisfy assumption (A5b) leads to the incorrect conclusion that treatment effects are heterogeneous over studies.

2.4 Treatment Assignment and Selection into Studies

We now consider the mechanisms by which subjects are assigned to treatments and selected into studies.

A standard condition for the identification of treatment effects in a single study is the assumption that treatment assignment is strongly ignorable given covariates (Rosenbaum and Rubin 1983), that is, within levels of the covariates, treatments are (or behave as if) randomly assigned to subjects with positive probability that may depend on subject covariates. We extend this assumption to the case of multiple studies:

A6. Strongly ignorable treatment assignment within studies given covariates: For all i, $\{Y_i(s,z) : s = 1,...,G, z = 1,...,L\} \underline{\perp\!\!\!\perp} Z_i \mid S_i = s, \mathbf{X}_i = \mathbf{x}$. For all i in study s and $z \in A_s$, $0 < \Pr(Z_i = z \mid S_i = s, \mathbf{X}_i = \mathbf{x}) < 1$.

Under assumptions (A1), (A2) and (A6), for any study s in which treatment z is administered, $F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y \mid Z = z, S = s, \mathbf{X} = \mathbf{x})$, that is, the observable distribution identifies the distribution of the potential outcomes in study s. If study s is a randomized study, assumption (A6) will be satisfied and treatment assignment will also be strongly ignorable: $\{Y_i(s,z) : s = 1,...,G, z = 1,...,L\} \underline{\perp\!\!\!\perp} Z_i \mid S_i = s$ and $0 < \Pr(Z_i = z \mid S_i = s) < 1$ for $z \in A_s$. In observational studies, treatment assignment will not generally be strongly ignorable, but may be strongly ignorable given covariates. However, if assumption (A6) is made in error, estimated treatment effects will generally be biased.

Next, we consider the mechanism by which subjects are selected into studies. If outcomes are unrelated to study selection, as would be the case if each study sampled the same population, potential outcomes are independent of studies: $\{Y_i(s,z) : s = 1, ..., G, z = 1, ..., L\} \underline{\perp\!\!\!\perp} S_i$. In general, however, different studies sample from different populations, in which case, even if $F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y(s',z) \mid S = s', \mathbf{X} = \mathbf{x})$, $F(y(s,z) \mid S = s) \neq F(y(s',z) \mid S = s')$. To account for this, we formalize the idea that differential selection on outcomes is due to observed covariates $\mathbf{X}_i$:

A7. Ignorable study selection, given covariates: For all i, $\{Y_i(s,z) : s = 1, ..., G, z = 1, ..., L\} \underline{\perp\!\!\!\perp} S_i \mid \mathbf{X}_i$.

Note that the same covariates $\mathbf{X}$ appear in both assumptions (A6) and (A7). In general, different covariates may account for treatment assignment and selection into studies. In this case, it will often be reasonable to pool the two sets of covariates and assume the combined set accounts for both treatment assignment and differential selection. In particular, if the studies are randomized and the covariates $\mathbf{X}$ account for selection into studies, as previously noted, assumption (A6) will also hold with respect to these covariates.

2.5 Combining Assumptions

We now examine the implications of the foregoing assumptions for conducting meta-analyses. Throughout, the extended stable unit treatment value assumption (A1), the sampling assumption (A2), and assumption (A6) that treatment assignment is strongly ignorable given covariates, are maintained. Under these assumptions, as noted earlier, $\mathrm{F}(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y \mid Z = z, S = s, \mathbf{X} = \mathbf{x})$; hence, for any study s in which treatments z and z' are administered, the treatment effect $H(F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}), F(y(s,z') \mid S = s, \mathbf{X} = \mathbf{x}))$ is identified.

If treatment z is administered in study s, and study selection is ignorable given covariates, for all s, s' and s'', $F(y(s,z) \mid S = s', \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid S = s'', \mathbf{X} = \mathbf{x})$, that is, for any value of the covariates $\mathbf{X}$, the potential outcomes have the same distribution in every study. If study selection is ignorable, that is, $\{Y_i(s,z) : s = 1, ..., G, z = 1, ..., L\} \underline{\parallel} S_i$, $F(y(s,z) \mid S = s') = F(y(s,z) \mid S = s'')$. Note that ignorable study selection does not imply study selection is ignorable given covariates, nor does ignorable study selection given covariates imply study selection is ignorable. However, if subjects in each of the studies are a sample from the same population, study selection is ignorable and also ignorable given covariates.

As previously noted, the response consistency assumptions cannot be tested directly. Similarly, assumption (A7) cannot be tested in the absence of additional assumptions. However, the weak response consistency assumption and assumption (A7) jointly imply the treatment effects $H(F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}), F(y(s,z') \mid S = s, \mathbf{X} = \mathbf{x}))$ are the same for all studies s. Further, the combination of these two assumptions is testable:

$$\begin{aligned} F(y \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid S = s, \mathbf{X} = \mathbf{x}) = F(y(s,z) \mid S = s', \mathbf{X} = \mathbf{x}) \\ = F(y(s',z) \mid S = s', \mathbf{X} = \mathbf{x}) = F(y \mid Z = z, S = s', \mathbf{X} = \mathbf{x}), \end{aligned} \quad (6)$$

that is, for every study s in which treatment z is administered, the distributions of the response, conditional on $\mathbf{X}$, are identical.

Thus, if (6) fails to hold, at least one of assumptions (A3b) or (A7) must be incorrect. In this case, if an investigator is willing to maintain the weak response consistency assumption, but not the selection assumption (A7), unobserved variables U related to the potential responses are differentially distributed across studies. On the other hand, if an investigator is willing to maintain the selection assumption (A7) but not the weak response consistency assumption, any between study heterogeneity in treatment effects stems from

unit heterogeneity in the potential outcomes, that is, subjects would have different values on the response variable in different studies.

We now consider the weak equivalence assumption (A5b), which is easily assessed for all studies in which both treatments z_1 and z_2 are administered 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 $F(y \mid S = s, Z = z_2, \mathbf{X} = \mathbf{x}) = F(y(s, z) \mid S = s, Z = z_2, \mathbf{X} = \mathbf{x})$. If the hypothesis of equality is accepted, this may suggest the extension of the weak equivalence hypothesis to studies where only one of the treatments is administered is plausible.

However, in meta-analyses where treatments z_1 and z_2 do not jointly appear in a study, as in two arm studies comparing these treatments with placebo z_0, the weak equivalence assumption is not directly testable. When it is assumed and it is also reasonable to make the weak response consistency assumption (A3b) and assumption (A7) that study selection is ignorable given covariates, weak equivalence can be 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 $F(y \mid S = s', Z = z_2, \mathbf{X} = \mathbf{x}) = F(y(s', z_2) \mid S = s', \mathbf{X} = \mathbf{x})$.

3 VIOXX AND CARDIOVASCULAR RISK: A META-ANALYSIS OF THE MERCK STUDIES

Vioxx is a COX-2 selective, non-steroidal 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 with 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 demonstrated 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.

The potential for COX-2 inhibitors to cause adverse cardiovascular events was recognized early in their development. By early 1998, scientists at Merck and Company Inc., where Vioxx was developed, were also aware of potential cardiotoxicity. Consequently, in 1998 Merck established and thereafter continued to use a standard operating procedure (SOP) that systematically monitored and adjudicated a wide array of thrombotic events.

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 (combining doses of 12.5 milligrams per day (mg/d), 25 mg/d and 50 mg/d in the treatment group) with subjects on placebo, naproxen and other NSAID's, finding Vioxx significantly increases the risk of MI. No study heterogeneity was found in the relative risk, nor did sub-group 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, valdecoxib), with placebo, finding that COX-2 inhibitors significantly increased the risk of experiencing serious (MI, stroke or vascular death) vascular events. 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. Zhang et al. (2006) conducted a study-level meta-analysis to assess the effects of Vioxx on renal events that included 114 randomized trials, finding Vioxx increases the risk of renal events.

Meta-analyses with individual patient 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 et al.

2002; Weir et al. 2003), concluding that Vioxx does 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 studies, all of which were completed before the withdrawal of Vioxx from the market, included in our analysis. As in Ross et al. (2009), all 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. Further, 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. 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. In total, there are 14,641 patients and 6,676 patient-years from 29 studies. However, the outcome of interest is sparse; there are 8 studies with no events in either arm and 9 studies with only one adverse event. The sparsity makes it difficult to analyze the data study by study; thus meta-analysis becomes crucial in 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 equation (2). Throughout, we make the extended SUTVA assumption (A1), the random sampling assumption (A2), and the assumption of strong ignorability given covariates (A6); in fact, as all studies are randomized, the stronger ignorability assumption $\{Y_i(s,z) : s = 1,...,G, z = 1,...,L\}, \mathbf{X}_i \underline{\|} Z_i \mid S_i = s$ holds. Under

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/d)	Duration (weeks)
010	Osteoarthritis	25/125	6
029	Osteoarthritis	12.5/25/50	6
033	Osteoarthritis	5/12.5/25	6
040	Osteoarthritis	12.5/25	6
044	Osteoarthritis	25/50	24
045	Osteoarthritis	25/50	24
058	Osteoarthritis	12.5/25	6
083	Osteoarthritis	25	64
085	Osteoarthritis	12.5	6
090	Osteoarthritis	12.5	6
112	Osteoarthritis	12.5/25	6
116	Osteoarthritis	25	6
136	Osteoarthritis	25	12
219	Osteoarthritis	12.5	6
220	Osteoarthritis	12.5	6
017	Rheumatoid Arthritis	125/175	6
068	Rheumatoid Arthritis	5/25/50	8
096	Rheumatoid Arthritis	12.5/25	12
097	Rheumatoid Arthritis	25/50	12
098	Rheumatoid Arthritis	50	12
103	Rheumatoid Arthritis	50	12
078	Alzheimer's Disease	25	208
091	Alzheimer's Disease	25	52
126	Alzheimer's Disease	25	52
118	Chronic Nonbacterial Prostatitis	25/50	6
120	Low Back Pain	25/50	4
121	Low Back Pain	25/50	4
125	Migraine Prophylaxis	25	12
129	Familial Adenomatous Polyposis	25	24

these assumptions,

$$\log(1 - F_{Y(s,z)}(t \mid S = s, \mathbf{X} = \mathbf{x}))/\log(1 - F_{Y(s,z')}(t \mid S = s, \mathbf{X} = \mathbf{x})) =$$
$$\log(1 - F_{Y}(t \mid Z = z, S = s, \mathbf{X} = \mathbf{x}))/\log(1 - F_{Y}(t \mid Z = z', S = s, \mathbf{X} = \mathbf{x})). \quad (7)$$

We also make the assumption that the censoring mechanism is non-informative. To estimate (2), we model the hazard $h(t, s, z \mid Z = z, S = s, \mathbf{X} = \mathbf{x})$ using the Cox proportional hazards model (Cox 1972); the effect (2) is then equal to the hazard ratio comparing treatments z and z' in the model. All models were estimated using the `survival` package (Therneau 2013) in `R` (version 3.0.0).

The weak response consistency assumption (see assumption (A3b)) is also made. If there were variation in the treatment, for example, a dose of 12.5 mg of Vioxx in study A is not equivalent to a dose of 12.5 mg in study B, or if the outcomes were coded differently in different studies, this assumption would be unreasonable. In this case, one might model the study-by-treatment interactions as random effects. However, as there is no reason to believe such variation is present and the CVT adverse events are coded by medical professionals, the weak consistency assumption seems reasonable. Under this assumption, any treatment by study interaction must be due to the differential selection of subjects into studies; in this case, we prefer to estimate interaction parameters. (Nevertheless, we performed the analysis using both fixed and random effects for the interactions, and the results are essentially the same.)

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; in no case did we reject this assumption at the .05 level. Thus, we do not fit 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 function for patients in study s with covariates $\mathbf{X}$ (which includes the individual level covariates $\mathbf{V}$) receiving treatment $Z = z$ is specified as:

$$\log h(t, s, z \mid Z = z, S = s, \mathbf{X} = \mathbf{x}) = \log h_0(t) + \alpha_s + \boldsymbol{v}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{\theta} + \boldsymbol{z}^T\boldsymbol{\Theta}\boldsymbol{v} + \eta_{sz}, \quad (8)$$

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 $\mathbf{V}$; the vector $\mathbf{X}$ also includes the study level variable "indications", which cannot be included in models incorporating the study parameters α_s. $\mathbf{Z}$ is a vector of indicator variables indexing the 5 dosage levels: $Z_1 = 1$ for receipt of a 12.5 mg/d dose, 0 otherwise, ..., $Z_4 = 1$ 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 $T = \{s, z$: treatment z is administered in study $s\}$, are the treatment by study interactions.

Under the assumption that responses (or treatment effects) are weakly consistent, the treatment by study interaction parameters η_{sz} are 0 if there is no differential selection into studies. Therefore, we compared the fit of model M2, in which the treatment effects do not depend on study, to that of model M1; the likelihood ratio test does not suggest that model M1 fits the data better (- 2 $\log \lambda = 38.4$, df = 40, p = .54, where λ is the likelihood ratio).

Model M3 is the special case of M2 with no treatment by covariate interaction. 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 is a further simplification of model M3. Here, the study parameters α_s are set to 0, and the study indicators replaced by a vector of indicator variables for the 7 study indications: Osteoarthritis, Rheumatoid Arthritis, Alzheimer's, Adenomatous Polyposis, Lower

Back Pain, Prostatitis, Migraine). The substitution of study indications for study does not result in a significant loss of information (-2 $\log \lambda = 27.8$, df = 22, p = .18); thus model M4 is preferred to model M3.

Model M4 is consistent with the weak consistency assumption and the assumption (A7) that the covariates $\mathbf{X}$ account for differential selection into studies, as these jointly imply the hazard does not depend on study. Further simplification is achieved in models M5i and M5. In model M5i, study indication is included, but the individual level covariates are not; however, model M5i fits substantially worse than M4 (-2 $\log \lambda = 58.41$, df = 2, p = .00). Furthermore, model M5, which includes the individual level covariates age and gender, but not study indication, is preferred to M4 (-2 $\log \lambda = 0.81$, df = 6, p = .99).

Further simplification would be possible if the differential selection assumption (A7) held unconditionally. In this case, model M6, in which the hazard depends only on treatment, would hold. However, the fit of this model is significantly worse than that of model M5 (-2 $\log \lambda = 75.24$, df = 2, p = .00).

Estimates of the treatment effects and 95% confidence intervals under model M5 are 1) 2.63 and (1.3, 5.3) for 12.5 mg/d, 2) 1.33 and (0.99, 1.78) for 25 mg/d, 3) 2.38 and (1.03, 5.5) for 50 mg/d , 4) 14 and (3.18, 61.63) for 125 mg/d.

Table 2: Models fit to the data. Covariates include age and gender. Treatment is an unordered categorical variable.

	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
M5i	treatment + indication
M6	treatment

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 mg/d, 25 mg/d and 50 mg/d arms (as more than 90% of the patient-weeks of observation in their data were contributed by subjects receiving the 25 mg/d dose), treating these doses as equivalent. Table 3 displays the number of patients and days-at-risk in the placebo and treatment arms for the data we analyzed. We now address the relationship between dosage and response.

Table 3: Total number of patients and days-at-risk under placebo and treatment arms.

	Placebo	Vioxx 12.5	Vioxx 25	Vioxx 50	Vioxx 125
patients	5451	2462	5181	1432	115
days-at-risk	1127971	122532	1057899	123931	4402

To begin, we examine a restricted version of model M5 (M5g) in which 12.5 mg/d, 25 mg/d and 50 mg/d doses are grouped into a low treatment level, with 125 mg/d 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 this model is not rejected by comparison with model M5 at the .05 level, (-2 $\log \lambda = 4.579$, df = 2, p = .10), a dose-response relationship is suggested. Further, the more restricted model M5rg that groups 12.5 mg/d, 25 mg/d, 50 mg/d and 125 mg/d into a single treatment level is rejected in comparison with model M5g (-2 $\log \lambda = 4.875$, 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 $\log \lambda = 6.156$, df = 3, p = .10) and model M5q is not preferred to M5ℓ (-2 $\log \lambda = 0.906$, 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 non-equivalent doses are combined, a researcher might be led to falsely conclude that responses are heterogeneous across studies. That occurs here. 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

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 ignorable given covariates. We assume the responses are consistent. While this assumption is not empirically testable 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, as here. Covariates are introduced into the analysis to account for possible differences in sub-population treatment effects and to account for differential selection into studies. The null hypothesis of no study by treatment interaction is not rejected at the .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 .05 level, implying the

response distributions for each treatment do not vary by study, conditional on covariates.

When the response consistency assumption holds and treatments have not been grouped (or the equivalence assumption holds), any heterogeneity in study level treatment effects is due to the differential distribution of covariates across studies. In this case, if individual patient level data are available and measured covariates account for the differential distribution of subjects across studies, hence 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 et al. 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 inconsistency, inconsistency nevertheless stems from

the sources of heterogeneity identified here, as homogeneity implies consistency. As Higgins et al. (2012) point out, when inconsistency is detected in a network meta-analysis, it is not entirely clear how the analysis should subsequently proceed. Although that is not of concern here, 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 inconsistency more generally when this phenomenon is encountered, and in future research we intend to examine this issue further.

References

Aitkin, M. (1999), "Meta-analysis by Random Effect Modelling in Generalized Linear Models," *Statistics in Medicine*, 18, 2343–2351.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Campbell, D. T., Stanley, J. C., and Gage, N. L. (1963), *Experimental and Quasi-experimental Designs for Research*, Houghton Mifflin Boston.

Covey, J. (2007), "A Meta-analysis of the Effects of Presenting Treatment Benefits in Different Formats," *Medical Decision Making*, 27, 638–654.

Cox, D. R. (1972), "Regression Models and Life-tables," *Journal of the Royal Statistical Society. Series B*, 187–220.

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

Greenland, S. (1994), "Invited Commentary: a Critical Look at Some Popular Meta-analytic Methods," *American Journal of Epidemiology*, 140, 290–296.

Higgins, J., Jackson, D., Barrett, J., Lu, G., Ades, A., and White, I. (2012), "Consistency and Inconsistency in Network Meta-analysis: Concepts and Models for Multi-arm Studies," *Research Synthesis Methods*, 3, 98–110.

Higgins, J., Thompson, S. G., and Spiegelhalter, D. J. (2009), "A Re-evaluation of Random-effects Meta-analysis," *Journal of the Royal Statistical Society: Series A*, 172, 137–159.

Higgins, J., Whitehead, A., Turner, R. M., Omar, R. Z., and Thompson, S. G. (2001), "Meta-

analysis of Continuous Outcome Data from Individual Patients," *Statistics in Medicine*, 20, 2219–2241.

Hudgens, M. G. and Halloran, M. E. (2008), "Toward Causal Inference with Interference," *Journal of the American Statistical Association*, 103.

Jüni, P., Nartey, L., Reichenbach, S., Sterchi, R., Dieppe, P. A., and Egger, M. (2004), "Risk of Cardiovascular Events and Rofecoxib: Cumulative Meta-analysis," *The Lancet*, 364, 2021–2029.

Kearney, P. M., Baigent, C., Godwin, J., Halls, H., Emberson, J. R., and 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, 1302–1308.

Konstam, M. A., Weir, M. R., Reicin, A., Shapiro, D., Sperling, R. S., Barr, E., and Gertz, B. J. (2001), "Cardiovascular Thrombotic Events in Controlled, Clinical Trials of Rofecoxib," *Circulation*, 104, 2280–2288.

Li, Y., Taylor, J. M., Elliott, M. R., and Sargent, D. J. (2011), "Causal Assessment of Surrogacy in a Meta-analysis of Colorectal Cancer Trials," *Biostatistics*, 12, 478–492.

Lumley, T. (2002), "Network Meta-Analysis for Indirect Treatment Comparisons," *Statistics in Medicine*, 21, 2313–2324.

Reicin, A. S., Shapiro, D., Sperling, R. S., Barr, E., Yu, Q., et al. (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, 204.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Ross, J. S., Madigan, D., Hill, K. P., Egilman, D. S., Wang, Y., and Krumholz, H. M. (2009), "Pooled Analysis of Rofecoxib Placebo-controlled Clinical Trial Data: Lessons for Postmarket Pharmaceutical Safety Surveillance," *Archives of Internal Medicine*, 169, 1976.

Rubin, D. B. (1980), "Comment on 'Randomization Analysis of Experimental Data: The Fisher Randomization Test,' by D. Basu," *Journal of the American Statistical Association*, 75, 591–593.

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, 1398–1407.

Therneau, T. M. (2013), *A Package for Survival Analysis in S*, package version 2.37-4.

Weir, M. R., Sperling, R. S., Reicin, A., and Gertz, B. J. (2003), "Selective COX-2 Inhibition and Cardiovascular Effects: A Review of the Rofecoxib Development Program," *The American Heart Journal*, 146, 591–604.

Zhang, J., Ding, E. L., and Song, Y. (2006), "Adverse Effects of Cyclooxygenase 2 Inhibitors on Renal and Arrhythmia Events," *JAMA: the Journal of the American Medical Association*, 296, 1619–1632.