The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 18

An Alternative to Pooling Kaplan-Meier Curves in Time-to-Event Meta-Analysis

Daniel B. Rubin, *Food and Drug Administration*

Recommended Citation:
Rubin, Daniel B. (2011) "An Alternative to Pooling Kaplan-Meier Curves in Time-to-Event Meta-Analysis," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 18.
DOI: 10.2202/1557-4679.1289

An Alternative to Pooling Kaplan-Meier Curves in Time-to-Event Meta-Analysis

Daniel B. Rubin

Abstract

A meta-analysis that uses individual-level data instead of study-level data is widely considered to be a gold standard approach, in part because it allows a time-to-event analysis. Unfortunately, with the common practice of presenting Kaplan-Meier survival curves after pooling subjects across randomized trials, using individual-level data can actually be a step backwards; a Simpson's paradox can occur in which pooling incorrectly reverses the direction of an association. We introduce a nonparametric procedure for synthesizing survival curves across studies that is designed to avoid this difficulty and preserve the integrity of randomization. The technique is based on a counterfactual formulation in which we ask what pooled survival curves would look like if all subjects in all studies had been assigned treatment, or if all subjects had been assigned to control arms. The method is related to a Kaplan-Meier adjustment proposed in 2005 by Xie and Liu to correct for confounding in nonrandomized studies, but is formulated for the meta-analysis setting. The procedure is discussed in the context of examining rosiglitazone and cardiovascular adverse events.

KEYWORDS: meta-analysis, survival analysis, Simpson’s paradox

Author Notes: This work only concerns the opinions of the author, and does not necessarily represent the views of the Food and Drug Administration.

1 Introduction

A common disclaimer when publishing a meta-analysis is that the investigation is limited because a time-to-event analysis cannot be performed without individual-level data. However, even with access to these data it is not always clear how studies should be combined; pooling subjects across trials can lead to biased conclusions, and hazard ratio modeling does not always describe how risk changes over time. Hence, we believe that time-to-event meta-analysis can benefit from new methodology.

1.1 Pooling Can Be Problematic

Simpson's paradox (Simpson, 1951) is the phenomenon in which within several studies or subgroups there can be an apparent association (e.g., treatment is harmful), but when data are pooled over all studies or subgroups the direction of association appears to reverse (e.g., treatment is helpful). The most common examples are observational studies where the paradox can be explained by confounding, but it is well-known that the phenomenon can also occur in a meta-analysis of randomized trials.

For example, Nissen and Wolski (2007) compared myocardial infarction rates of subjects randomized to diabetes drug rosiglitazone (Avandia) or comparator drugs in a meta-analysis of approximately 28,000 subjects in 42 trials. Using study-level data and a fixed effects model they estimated an odds ratio for myocardial infarction of 1.43 and found this to be statistically significant. However, Bracken (2007) pointed that when the 42 studies were pooled there was a Simpson's paradox in which the direction of association reversed. Event rates in pooled rosiglitazone and control arms were 5.5/1000 and 5.9/1000, and using pooled rates gave odds ratio estimate $0.94 < 1$.

Rücker and Schumacher (2008) summarize why this paradox can occur, and note that a meta-analysis can be susceptible when treatment assignment probabilities differ between trials. Unequal assignment probabilities are not typically part of trial designs, but can appear when combining studies in a meta-analysis because is common to collapse all dose or comparator arms within a study into one treatment group and one control group.

This fact may raise questions about a common practice in meta-analysis with time-to-event data that has been used to assess cardiovascular risks for rosiglitazone and a drug in the same class: pooling treatment and control groups across studies and computing Kaplan-Meier curves (Lincoff et al.,

2007; FDA, 2007; Cobitz et al., 2008). Key questions arising in 2007 and 2010 FDA advisory committee meetings on rosiglitazone were whether the signal was driven by events in the first six months after treatment assignment, and whether any excess risk disappeared or reversed direction upon long-term follow-up. Answering such questions clearly requires time-to-event analysis.

In recognition of the problems associated with pooling survival data over studies and the lack of alternative methodologies, the Cochrane Collaboration (Higgins and Green, 2009) has stated that

> Kaplan Meier plots for all pooled participants across trials in a meta-analysis have previously been presented in medical journals. This practice breaks with the principle of comparing like with like. For this reason, until further discussions have taken place the Statistical Methods Group is unable to recommend inclusion of such plots in Cochrane reviews.

As a hypothetical example illustrating the problem of pooling, suppose two studies are conducted, each with 1,000 total patients. In the first study, subjects assigned to control survive for three years on average, following an exponential distribution. The treatment is inferior to the control drug and cuts survival in half. In this first study 750 subjects are randomly assigned treatment, and the other 250 to the control. Subjects are followed for up to two years before being censored.

Subjects enrolled in the second hypothetical study are much sicker, and patients assigned to control only survive for six months on average, with survival again following an exponential distribution. The treatment again halves the survival time, so treated subjects live on average for three months according to the exponential distribution. Subjects are again followed for up to two years but assignment probabilities are now reversed, so only 250 of the 1,000 subjects are given the treatment.

After simulating the two trials, Figure 1 shows estimated survival curves after pooling studies. Even though treatment is always harmful, pooling leads to the incorrect conclusion that it increases survival. Simpson's paradox occurs because the pooled treatment arm has more of the healthier subjects from the first study. The finding was not due to simulation error: over 200 simulated meta-analyses the estimated six month survival rate in the control arm averaged 0.49 +/- 0.01 SD, while the estimated treatment survival rate at six months averaged a larger 0.57 +/- 0.01 SD.

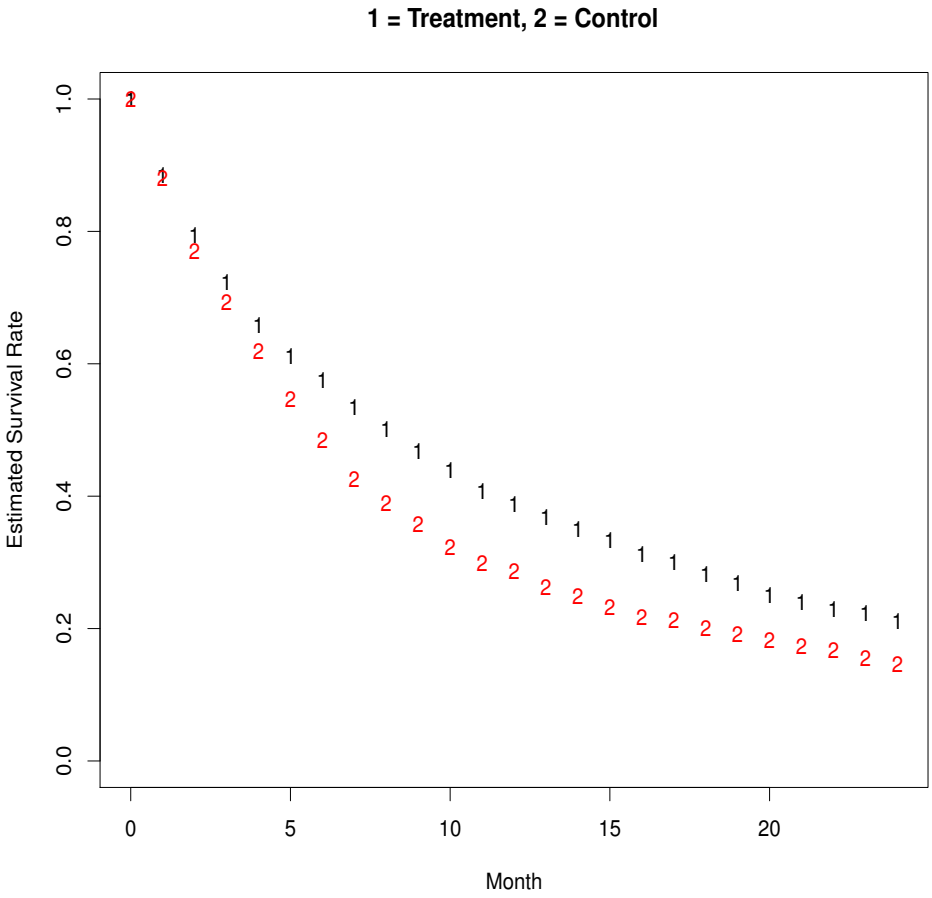

Figure 1: Estimated survival rates at every month between 0-2 years when pooling subjects across the two simulated studies. Even though treatment is always harmful and the studies are randomized, there is a Simpson's paradox in which pooling makes the treatment appear beneficial.

A possible counterargument to this example is that there is heterogeneity in the two studies, so they are not inherently combinable. In general we agree that studies in a meta-analysis should examine similar interventions in similar patient populations. However, heterogeneity can be difficult to detect, particularly with rare events, although it could probably be found in this example. Nevertheless, if diagnostics do not preclude meta-analysis then the example shows that pooling can be perilous.

1.2 Hazard Ratios Might Not Tell the Whole Story

Another approach to meta-analysis with survival data is to model hazard ratios. For instance, with individual-level data one could fit a proportional hazards model and stratify by study, which allows baseline hazard functions to differ between trials. Additionally, one could combine estimated hazard ratios across studies. The combination would typically involve a weighted average using a fixed effects model or a random effects model. While we do not dispute that hazard ratios can often be very informative, survival probabilities at different times can have a clearer interpretation in some contexts. For instance, a clinician or patient might be interested in the overall event rate at six months under treatment and control. Likewise, traditional hazard ratio modeling cannot describe "the time course of risks" that has been sought in the case of rosiglitazone (Psaty and Furberg, 2007), because a proportional hazards model assumes the hazard ratio is constant over time, and estimation only depends on failure and censoring times through their rank order.

2 Counterfactual View of Randomized Trials

We approach randomized clinical trials through the lens of Neyman's (1923) potential outcome framework for experiments. To clarify our conceptual view of the problem we start by considering trials with binary outcomes before moving to time-to-event data. In this setting our procedure reduces to the "SS-based method" [study size-based method] of analysis discussed by Chuang-Stein and Beltangady (2010). The point is not to promote this method over more traditional fixed effects models or random effects models, but to introduce it because it more naturally generalizes to synthesizing survival curves.

DOI: 10.2202/1557-4679.1289

For the binary outcome problem, we consider the ideal but unavailable full data for k two-arm studies in a meta-analysis to be

$$\{w_i,\ y_i^T,\ y_i^C\}_{i=1}^n.$$

- Each (w_i, y_i^T, y_i^C) represents data for a single subject. These are considered fixed and deterministic, and are subject-level parameters.
- $w_i \in \{1, ..., k\}$ represents which of the k studies contains individual i.
- $y_i^T \in \{0, 1\}$ is a counterfactual outcome for individual i. If the subject is assigned to treatment, this is the binary outcome that will be seen.
- $y_i^C \in \{0, 1\}$ is another counterfactual outcome for individual i. If subject i is assigned to control, this will be the outcome.

The full data are unavailable because counterfactual outcomes y_i^T and y_i^C can never both be observed, since each subject can only be assigned to one of the two treatment arms. If subject i is assigned to treatment then only y_i^T will be observed, while under assignment to control only y_i^C will be seen. Thus, the individual-level meta-analysis observable data are

$$\{w_i,\ X_i,\ Y_i = y_i^T X_i + y_i^C(1 - X_i)\}_{i=1}^n.$$

- Random variable X_i is an binary indicator of treatment assignment for individual i, where $X_i = 1$ represents assignment to treatment and $X_i = 0$ represents assignment to control.
- Random variable Y_i is the observed outcome for subject i, which is y_i^T under assignment to treatment and y_i^C under assignment to control.

In randomized trials, treatment indicator X_i is randomly set to zero or one. Treatment assignment probabilities can differ across studies, but we assume that assignments made in different studies are independent of each other. Letting $n_1, ..., n_k$ denote the total number of subjects in each study, we assume a fixed number $m_1, ..., m_k$ of subjects are assigned treatment in each one, with $0 < m_l < n_l$. Thus, $\{X_i\}_{i=1}^n$ are not necessarily independent or identically distributed. For each individual we can also define the probability of assignment to treatment as

$$\pi_i = P(X_i = 1) = \sum_{l=1}^{k} 1(w_i = l)\frac{m_l}{n_l}\ .$$

In this statistical model, the only randomness is due to the random assignment of each subject to one of two arms (Freedman, 2008). The approach recognizes that subjects in clinical trials (and the trials themselves) are not usually randomly sampled from any well-defined population or superpopulation, but instead constitute convenience samples. Generalization to subjects outside the studies is not incorporated into the statistical modeling.

2.1 Parameters of Interest with Binary Outcomes

If sponsors of the clinical trials had only been interested in studying subjects given the treatment, then all subjects in all studies could have been assigned to treatment. In that case, the observed data would be $\{w_i, y_i^T\}_{i=1}^n$. Unless some of the recruited subjects were considered more representative of a hypothetical target population than others, a natural summary measure of the event rate for subjects assigned treatment would be the empirical rate

$$y^T = n^{-1} \sum_{i=1}^{n} y_i^T.$$

Similarly, if sponsors were only interested in studying subjects given controls, all subjects in all studies could have been assigned to the control arm. With observed data $\{w_i, y_i^C\}_{i=1}^n$, a natural summary measure for outcomes of controls would be the empirical rate

$$y^C = n^{-1} \sum_{i=1}^{n} y_i^C.$$

In clinical trials, neither y^T or y^C will be available from the observed data, because treatment assignment probabilities are not set to zero or one. Thus, we consider these as summary parameters that would only be known to an oracle. To summarize the differential effect of treatment we could also, for example, attempt to estimate differences or ratios between y^T and y^C.

DOI: 10.2202/1557-4679.1289

2.2 Estimation

Recalling that $\pi_i = P(X_i = 1)$, the unavailable summary parameters can be estimated as follows:

$$Y^T = n^{-1} \sum_{i=1}^{n} \frac{X_i}{\pi_i} Y_i,$$
$$Y^C = n^{-1} \sum_{i=1}^{n} \frac{1 - X_i}{1 - \pi_i} Y_i.$$

As mentioned earlier, these reduce to the SS estimators discussed by Chuang-Stein and Beltangady (2010). The estimators are based on the inverse probability of treatment weighting approach (Robins et al., 1994), which generalizes a weighting scheme in survey sampling due to Horvitz and Thompson (1951). The idea in estimating the response rate y^T is to average responses of those who received treatment but give more weight to those with a lower probability of being treated, because they are more representative of the unobserved counterfactual responses for subjects in the control arm. It is simple to check that the estimators are unbiased.

Lemma 1. *$E[Y^T] = y^T$ and $E[Y^C] = y^C$.*

Proof. Note that $X_i Y_i = X_i y_i^T$ because $Y_i = y_i^T X_i + y_i^X (1 - X_i)$ and $X_i(1 - X_i) = 0$. Similarly, $(1 - X_i)Y_i = (1 - X_i)y_i^C$. Hence, term i in the two respective estimators is $(X_i/\pi_i)y_i^T$ and $(1 - X_i)/(1 - \pi_i)y_i^C$, which are unbiased for y_i^T and y_i^C. □

2.3 Avoiding Simpson's Paradox

Note that the SS estimators Y^T and Y^C can be computed without access to individual-level data. Let P_l^T and P_l^C denote observed event rates for treatment and control subjects in study l. It is easy to verify that these are unbiased for p_l^T and p_l^C, the event rates in study l under counterfactual assignment of all subjects to treatment or control. Recalling that n is the total number of subjects and n_l is the total number of subjects in study l, the estimators can be written as $Y^T = \sum_{l=1}^{k} \frac{n_l}{n} P_l^T$ and $Y^C = \sum_{l=1}^{k} \frac{n_l}{n} P_l^C$. These are weighted averages of study-level event rates for treatment and control, where the trial weight is simply the total number of subjects in the

trial, over both arms. Similarly, the desired parameters $y^T = \sum_{l=1}^{k} \frac{n_l}{n} p_l^T$ and $y^C = \sum_{l=1}^{k} \frac{n_l}{n} p_l^C$ are weighted averages of study-level parameters.

With this representation, it becomes clear why the SS method is immune to bias caused by unequal treatment assignment probabilities across studies. If treatment leads to higher event rates than the control in study l then risk difference $p_l^T - p_l^C$ should be positive. If this is true for all studies, then $y^T - y^C$ should also be positive as it is a weighted average of these terms. Consequently, if y^T and y^C are well-estimated then we will avoid any biased Simpson's paradox conclusions suggested by pooling.

2.4 Generalizing the Fixed Effects Model

A fixed effects model allows the study-level event rates (p_l^T, p_l^C) to differ from trial to trial, but assumes they live on a level set where a treatment effect such as a risk difference, risk ratio, or odds ratio is constant. If there is a constant study-level risk difference $\mu = p_l^T - p_l^C$ then the SS method estimates this, as $y^T - y^C = \sum_{l=1}^{k} \frac{n_l}{n}(p_l^T - p_l^C) = \sum_{l=1}^{k} \frac{n_l}{n}\mu = \mu$. Similarly, if there is a constant risk ratio $\mu = p_l^T / p_l^C$ then the SS method estimates it because $y^T / y^C = \sum_{l=1}^{k} \frac{n_l}{n} p_l^T / \sum_{l=1}^{k} \frac{n_l}{n} p_l^C = \sum_{l=1}^{k} \frac{n_l}{n} p_l^C \mu / \sum_{l=1}^{k} \frac{n_l}{n} p_l^C = \mu$. The main drawback of using SS method for estimation is a loss of efficiency relative to using Mantel-Haenszel weights. Unlike the risk difference or ratio, if there is a constant trial-level odds ratio then it may not equal the odds ratio formed from y^T and y^C, because the odds ratio is not "collapsible" (Ducharme and Lepage, 1986). However, for rare events as with rosiglitazone, the odds ratio will roughly equal the relative risk, so the SS method will still estimate the same parameter as a well-posed fixed effects model.

2.5 Application to Rosiglitazone

The safety of rosiglitazone is beyond our scope, so results mentioned here are to illustrate methodology. We do not mean to endorse any position on the safety profile of the drug or the combinability of the studies.

When applying the SS method to the aforementioned meta-analysis data for myocardial infarctions given by Nissen and Wolski (2007), which can be done with study-level data, we obtain $Y^T = 0.0067$ and $Y^C = 0.0048$. That is, we estimate that if all 28,000 subjects in all 42 studies had been assigned rosiglitazone we would have seen event rate 6.7/1000, and analogously, event

DOI: 10.2202/1557-4679.1289

rate 4.8/1000 if all subjects had been assigned to controls. The estimated odds ratio from (unrounded) Y^T and Y^C is 1.41. This is almost identical to the published fixed effects result of 1.43. We also tested the strong null hypothesis of no treatment effect (i.e., $y_i^T = y_i^C$) by randomly permuting treatment assignment labels in all trials 100,000 times and recomputing the SS odds ratio estimator in each permuted meta-analysis dataset to form a null distribution (Rosenbaum, 2002). This gave a two-sided $p = 0.02$, similar to Nissen and Wolski's $p = 0.03$. Nissen and Wolski also examined rates of death from cardiovascular causes. Their fixed effect model gave odds ratio estimate 1.64 ($p = 0.06$). The SS method yields $Y^T = 2.6/1000$ and $Y^C = 1.5/1000$ when estimating what the cardiovascular death rates would have been if all subjects had respectively been assigned to treatment or control, and odds ratio estimate 1.73 ($p = 0.02$).

An interesting difference between the SS method and standard fixed effect models is that the former does not exclude studies with zero events, which was controversial in the rosiglitazone meta-analysis (Diamond et al., 2007).

In summary, we find that the SS method avoids the Simpson's paradox induced by pooling and replicates a well-known result, which gives reassurance before generalizing to survival analysis.

3 Time-to-Event Meta-Analysis

Taking a similar counterfactual view as in the previous section, with time-to-event outcomes that are possibly right censored we define the unavailable full potential outcome data as

$$\{w_i,\ \delta_i^T,\ y_i^T,\ \delta_i^C,\ y_i^C\}_{i=1}^n.$$

- $w_i \in \{1, ..., k\}$ again indicates the study containing individual i. The full data $(w_i, \delta_i^T, y_i^T, \delta_i^C, y_i^C)$ for subject i are again considered deterministic subject-level parameters.

- δ_i^T is a counterfactual indicator of censoring. $\delta_i^T = 0$ indicates that the subject will be lost to follow-up if assigned treatment, while $\delta_i^T = 1$ indicates that failure will be observed under assignment to treatment.

- If subject i is assigned to treatment, $y_i^T \geq 0$ will be the time to either failure or loss to follow-up.

- $\delta_i^C \in \{0, 1\}$ is a counterfactual indicator of censoring when subject i is assigned to control.

- $y_i^C \geq 0$ is a counterfactual time to either failure or loss to follow-up when subject i is assigned to control.

If subject i is assigned to treatment then (δ_i^T, y_i^T) will be observed, and under assignment to control we will instead see (δ_i^C, y_i^C). As each subject can only be assigned to one of two arms, the observable individual-level data are

$$\{w_i,\ X_i,\ \Delta_i = \delta_i^T X_i + \delta_i^C (1 - X_i),\ Y_i = y_i^T X_i + y_i^C (1 - X_i)\}_{i=1}^n.$$

- Random variable X_i is again an indicator of treatment assignment for individual i, where $X_i = 1$ represents assignment to treatment and $X_i = 0$ represents assignment to control. As in the previous section, we can define $\pi_i = P(X_i = 1)$, determined from the number of subjects $m_1, ..., m_k$ assigned treatment in each study and the total number of subjects $n_1, ..., n_k$ in each study.

- Random variable Δ_i indicates whether the failure time for subject i is observed ($\Delta_i = 1$) or is censored ($\Delta_i = 0$).

- Random variable Y_i is the time to either failure or loss to follow-up for subject i.

Like our statistical model of meta-analysis with binary outcomes, the only stochastic component is the randomness induced by random assignment of subjects to treatment arms.

3.1 Survival Curves of Interest

In summarizing the survival distribution of subjects assigned to treatment, consider what would be done by an oracle who could redo all studies and assign all subjects to treatment. With time-to-event data $\{\delta_i^T, y_i^T\}_{i=1}^n$, a natural way to summarize survival would be through the Kaplan-Meier curve.

Likewise, when interested in the survival distribution of subjects assigned to control, an oracle could repeat all studies and assign every subject to the control arm. With time-to-event data $\{\delta_i^C, y_i^C\}_{i=1}^n$ for all subjects under control, the survival distribution could also be summarized through a Kaplan-Meier curve. Thus, we define our desired parameters as follows:

- f^T is the cumulative distribution function resulting from applying the Kaplan-Meier estimator to unavailable counterfactual data $\{\delta_i^T, y_i^T\}_{i=1}^n$.
- f^C is the cumulative distribution function resulting from applying the Kaplan-Meier estimator to unavailable counterfactual data $\{\delta_i^C, y_i^C\}_{i=1}^n$.

Note that the two unknown curves are considered parameters, not estimators, even though the Kaplan-Meier method is an estimation technique. This is similar to our approach with binary outcomes, in that we consider each subject to have deterministic data after assignment to treatment or control, and then define two parameters of interest as summary statistics that could be computed by an oracle when (a) all subjects in all studies are assigned to treatment; (b) all subjects in all studies are assigned to control. Because we work in a finite population urn model, another atypical feature of our formulation is that the curves of interest f^T and f^C can depend on the number of subjects and the number of studies.

3.2 Simpson's Paradox and Estimator Motivation

Of course, even an oracle's synthesized Kaplan-Meier curves might not always be easy to interpret. Consider the counterfactual time y_i^T for subject i to be the minimum of a failure time r_i^T and follow-up time z_i^T. It is the distribution of failure times $\{r_i^T\}_{i=1}^n$ that should truly be of interest rather than a curve involving the censoring mechanism. The oracle curve f^T will only approximate the distribution function $s \to n^{-1}\sum_{i=1}^n 1(r_i^T \leq s)$ of failure times if censoring is not very informative for failure, meaning $n^{-1}\sum_{i=1}^n 1(r_i^T \leq u, z_i^T \leq v) \approx n^{-1}\sum_{i=1}^n 1(r_i^T \leq u)1(z_i^T \leq v)$. For example, the survival curves could be misleading if trials with shorter follow-up times have healthier subjects than trials with longer follow-up periods.

Another potential censoring issue in the meta-analyses of rosiglitazone was that safety endpoints such as myocardial infarction did not always include death as a failure event. To illustrate the potential problem, note that if a hypothetical drug reduces fatal stroke then it might appear to increase the risk of myocardial infarction because sicker subjects are living longer and being followed longer. The simplest way to ensure that a treatment effect reflects a clinically meaningful signal rather than a drug-induced difference in follow-up times is to use composite failure events that include mortality, such as {myocardial infarction or death}.

Additionally, to interpret the oracle Kaplan-Meier curves the combined studies should ideally have similar endpoints, patient populations, and comparators, reflective of trials measuring similar interventions in similar target populations. In homogeneous trials we would expect the study-level failure time distributions $s \to n_l^{-1} \sum_{i=1}^{n} 1(w_i = l, r_i^T \leq s)$ to be approximately equal for studies $l = 1, ..., k$, and likewise for counterfactual failure times corresponding to the control arm.

Because we only recommend combining homogeneous trials with little informative censoring, it is natural to ask why we are concerned about pooling. After all, pooling may be valid under such conditions. Our response is that we desire a method that is robust in the sense that it (i) gives the appropriate survival curves if failure time distributions are indeed constant over trials; (ii) preserves randomization protection even if there is heterogeneity, so that relative to simple pooling the resulting curves may be better (albeit crude) reflections of the time course of risk.

Estimators of the desired curves f^T and f^C are immune to the usual cause of such confounding-by-trial bias and the consequent Simpson's paradox: unequal treatment assignment probabilities over the different studies. In our counterfactual model the desired curves f^T and f^C do not depend on the assignment probabilities. Even if some trials have much lower event rates than others, our method will attempt to estimate the same curves whether subjects in these trials are disproportionately assigned to treatment or whether all studies use 1:1 randomization.

In well-designed studies there will be no ascertainment bias, so the follow-up time for any subject will be the same under assignment to treatment or control (i.e., $z_i^T = z_i^C$). In this case any separation between survival curves f^T and f^C can be causally attributed to an effect of the treatment intervention on failure times $\{r_i^T\}_{i=1}^n$ and $\{r_i^C\}_{i=1}^n$, and Simpson's paradox will be avoided in the sense that $f^T - f^C$ will be nonnegative if $r_i^T \leq r_i^C$ for all subjects.

3.3 Re-Expressing the Desired Survival Curves

We will express the desired curves in a way that facilitates estimation. We begin with four initial functions. At time s, they are determined by the number of subjects who (counterfactually) are no longer in the risk set and the number who (counterfactually) have already had failure events:

DOI: 10.2202/1557-4679.1289

$$a^T(s) = n^{-1} \sum_{i=1}^{n} 1(y_i^T \leq s),$$
$$b^T(s) = n^{-1} \sum_{i=1}^{n} 1(\delta_i^T = 1,\ y_i^T \leq s),$$

and

$$a^C(s) = n^{-1} \sum_{i=1}^{n} 1(y_i^C \leq s),$$
$$b^C(s) = n^{-1} \sum_{i=1}^{n} 1(\delta_i^C = 1,\ y_i^C \leq s).$$

Next, use these functions to define λ^T and λ^C as the Nelson-Aalen cumulative hazard functions applied to the unavailable counterfactual time-to-event data $\{\delta_i^T, y_i^T\}_{i=1}^n$ and $\{\delta_i^C, y_i^C\}_{i=1}^n$ in the treatment and control arms:

$$\lambda^T(s) = \int_{[0,\ s]} \frac{db^T(u)}{1 - a^T(u-)},$$
$$\lambda^C(s) = \int_{[0,\ s]} \frac{db^C(u)}{1 - a^C(u-)}.$$

These integrals here reduce to finite sums. To be explicit, let $\{u_j^T\}_{j=1}^p$ denote the distinct sorted values of $\{y_i^T\}_{i=1}^n$ such that $\delta_i^T = 1$, and likewise let $\{u_j^C\}_{j=1}^q$ denote the distinct sorted values of $\{y_i^C\}_{i=1}^n$ such that $\delta_i^C = 1$. As bookkeeping, define $u_0^T = u_0^C = -\infty$ and $b^T(u_0^T) = b^C(u_0^C) = 0$. Noting that $a^T(u-)$ and $a^C(u-)$ are $n^{-1} \sum_{i=1}^n 1(y_i^T < u)$ and $n^{-1} \sum_{i=1}^n 1(y_i^C < u)$, the cumulative hazard functions are

$$\lambda^T(s) = \sum_{j=1}^{p} 1(u_j^T \leq s) \frac{b^T(u_j^T) - b^T(u_{j-1}^T)}{1 - a^T(u_j^T-)},$$
$$\lambda^C(s) = \sum_{j=1}^{q} 1(u_j^C \leq s) \frac{b^C(u_j^C) - b^T(u_{j-1}^C)}{1 - a^C(u_j^C-)}.$$

Finally, the Kaplan-Meier CDFs f^T and f^C can be defined through applying the product integral operator π to the cumulative hazard functions to obtain

$$f^T(s) = 1 - \prod_0^s \left(1 - d\lambda^T(u)\right),$$
$$f^C(s) = 1 - \prod_0^s \left(1 - d\lambda^C(u)\right).$$

We refer to Gill and Johansen (1990) for details on why this gives the Kaplan-Meier curves and for a summary of the product integral. Product integration is a generalization of ordinary multiplication that can be applied to continuous or discrete time, just as a regular integration is a generalization of summation. We will not formally define the general operator, because the product integrals needed to define the two desired Kaplan-Meier curves reduce to ordinary products with our discrete counterfactual data. To be explicit, consider the times $\{u_j^T\}_{j=1}^p$ and $\{u_j^C\}_{j=1}^q$ previously defined, and as bookkeeping define $\lambda^T(u_0^T) = \lambda^C(u_0^C) = 0$. The desired cumulative distribution functions can then be written as

$$f^T(s) = 1 - \prod_{j=1}^{p} \left(1 - (\lambda^T(u_j^T) - \lambda^T(u_{j-1}^T))\right)^{1(u_j^T \le s)},$$
$$f^C(s) = 1 - \prod_{j=1}^{q} \left(1 - (\lambda^C(u_j^C) - \lambda^C(u_{j-1}^C))\right)^{1(u_j^C \le s)}.$$

3.4 Estimation

Starting with (a^T, b^T) and (a^C, b^C) defined from the unavailable full counterfactual data, we mapped these to the desired f^T and f^C. Intuitively, if we could estimate the initial (a^T, b^T) and (a^C, b^C), we could estimate f^T and f^C by applying the same mapping. This is the route we take. We begin with

$$A^T(s) = n^{-1} \sum_{i=1}^{n} \frac{X_i}{\pi_i} 1(Y_i \le s)$$
$$B^T(s) = n^{-1} \sum_{i=1}^{n} \frac{X_i}{\pi_i} 1(\Delta_i = 1,\ Y_i \le s),$$

DOI: 10.2202/1557-4679.1289

and

$$A^C(s) = n^{-1} \sum_{i=1}^{n} \frac{1 - X_i}{1 - \pi_i} 1(Y_i \leq s),$$

$$B^C(s) = n^{-1} \sum_{i=1}^{n} \frac{1 - X_i}{1 - \pi_i} 1(\Delta_i = 1,\ Y_i \leq s).$$

Following the reasoning in Lemma 1, it is simple to check that all four of these random functions are unbiased at any time s.

The next step is estimating cumulative hazard functions for each arm. Replacing the four unknown functions with the estimated functions we just defined, we obtain

$$\Lambda^T(s) = \int_{[0,\, s]} \frac{dB^T(u)}{1 - A^T(u-)},$$

$$\Lambda^C(s) = \int_{[0,\, s]} \frac{dB^C(u)}{1 - A^C(u-)}.$$

As before when defining the Nelson-Aalen cumulative hazards on the unavailable full data, these integrals become finite sums. Let $\{U_j^T\}_{j=1}^P$ denote the distinct sorted values of times $\{Y_i\}_{i=1}^n$ such that $X_i = 1$ and $\Delta_i = 1$, and likewise let $\{U_j^C\}_{j=1}^Q$ denote the distinct sorted values of $\{Y_i\}_{i=1}^n$ such that $X_i = 0$ and $\Delta_i = 1$. As bookkeeping, define $U_0^T = U_0^C = -\infty$ and $B^T(U_0^T) = B^C(U_0^C) = 0$. Noting that $A^T(u-)$ and $A^C(u-)$ can be evaluated as $n^{-1} \sum_{i=1}^n (X_i/\pi_i) 1(Y_i < u)$ and $n^{-1} \sum_{i=1}^n (1 - X_i)/(1 - \pi_i) 1(Y_i < u)$, the estimated cumulative hazard functions are

$$\Lambda^T(s) = \sum_{j=1}^{P} 1(U_j^T \leq s) \frac{B^T(U_j^T) - B^T(U_{j-1}^T)}{1 - A^T(U_j^T-)},$$

$$\Lambda^C(s) = \sum_{j=1}^{Q} 1(U_j^C \leq s) \frac{B^C(U_j^C) - B^T(U_{j-1}^C)}{1 - A^C(U_j^C-)}.$$

The last step is mapping the estimated cumulative hazard functions into estimated CDFs. Following the last section, estimators can be formed through applying product integrals to obtain

$$F^T(s) = 1 - \prod_0^s \left(1 - d\Lambda^T(u)\right),$$

$$F^C(s) = 1 - \prod_0^s \left(1 - d\Lambda^C(u)\right).$$

To be explicit, first evaluate Λ^T and Λ^C at the times $\{U_j^T\}_{j=1}^P$ and $\{U_j^C\}_{j=1}^Q$, and as bookkeeping define $\Lambda^T(U_0^T) = \Lambda^C(U_0^C) = 0$. Then the estimated CDFs are the finite products

$$F^T(s) = 1 - \prod_{j=1}^{P} \left(1 - (\Lambda^T(U_j^T) - \Lambda^T(U_{j-1}^T)\right)^{1(U_j^T \leq s)},$$

$$F^C(s) = 1 - \prod_{j=1}^{Q} \left(1 - (\Lambda^C(U_j^C) - \Lambda^C(U_{j-1}^C)\right)^{1(U_j^C \leq s)}.$$

Thus, we can estimate f^T with F^T by applying the same mapping to functions (A^T, B^T) that was applied to (a^T, b^T) in defining the curve of interest, and likewise for f^C. The hope is that F^T and F^C will be good approximations because in large samples (A^T, B^T, A^C, B^C) will be close to (a^T, b^T, a^C, b^C), and we will formalize this notion through asymptotics.

3.5 Large Sample Consistency

In the lemma below we show that our method is consistent, in that the estimated survival curves will approximate the oracle curves in large samples. That is, we define "consistency" with respect to an oracle, and not in an absolute sense. We consider a sequence of meta-analyses, and to be technical these should be arranged in a triangular array. Meta-analysis j in the sequence should have k_j studies and n_j total subjects, the full data should be written as $\{w_{i,j}, \delta_{i,j}^T, y_{i,j}^T, \delta_{i,j}^C, y_{i,j}^C\}_{i=1}^{n_j}$, and the observable data should be written as $\{w_{i,j}, X_{i,j}, \Delta_{i,j}, Y_{i,j}\}_{i=1}^{n_j}$. The number of subjects in each study should be denoted by $n_{1,j}, ..., n_{k_j,j}$ and the number assigned to treatment should be denoted $m_{1,j}, ..., m_{k_j,j}$. However, we suppress notation for convenience and omit double-indexing of the array in the statement below. The proof is deferred to the appendix.

Lemma 2. *Consider a sequence of meta-analyses such that*

$$(a^T(s), b^T(s)) \to (a_\infty^T(s), b_\infty^T(s)) \text{ and } (a^T(s-), b^T(s-)) \to (a_\infty^T(s-), b_\infty^T(s-))$$

for each $s \in [0, \tau]$*, where* $a_\infty^T(\tau-) < 1$*. Suppose also that*

$$(a^C(s), b^C(s)) \to (a_\infty^C(s), b_\infty^C(s)) \text{ and } (a^C(s-), b^C(s-)) \to (a_\infty^C(s-), b_\infty^C(s-))$$

for each $s \in [0, \tau]$*, where* $a^C_\infty(\tau-) < 1$*. Recalling that* k *and* n *represent the number of studies and number of subjects, and letting* $\omega_l = n_l/n$ *and* $\gamma_l = m_l/n_l$ *represent the proportion of subjects in study* l *and the treatment assignment probability in study* l*, assume that*

$$n^{-1} \left(\sum_{l=1}^{k} \omega_l/\gamma_l \right) \to 0, \tag{1}$$

$$n^{-1} \left(\sum_{l=1}^{k} \omega_l/(1-\gamma_l) \right) \to 0. \tag{2}$$

Then the estimated cumulative distribution functions are uniformly consistent for their target parameters on $[0, \tau]$*, in that*

$$sup_{s\in[0,\ \tau]}|F^T(s) - f^T(s)| \to_p 0,$$
$$sup_{s\in[0,\ \tau]}|F^C(s) - f^C(s)| \to_p 0.$$

The restriction to estimation of survival on $[0,\ \tau]$ is standard, as a survival curve cannot be approximated after times at which all subjects are censored.

For (1) and (2) to be small, as is needed for the lemma, assignment probabilities should be kept away from zero and one, because otherwise either the treatment distribution or control distribution could not be estimated in some studies. Note that the two terms are made small if either there are a fixed number of studies and the total number of subjects is large, or if there are many studies and none contain a large proportion of total subjects.

3.6 Simulated Example

We return to the two simulated studies considered in Section 1.1. Recall that treatment halved survival in both studies, and that there was a Simpson's paradox in which Kaplan-Meier curves from subjects pooled over two studies incorrectly suggested that treatment was beneficial. Figure 2 shows our newly-defined survival estimators $1 - F^T$ and $1 - F^C$ applied to the previously generated time-to-event data. The important thing to notice is that we now at least avoid Simpson's paradox; unlike ordinary pooling the curves identify that treatment is harmful at every follow-up time before two years. This correct finding is not due to simulation error. For example, over 200 simulated meta-analyses the estimated six month survival rate in the control arm averaged 0.61 +/- 0.01 SD and the estimated six month treatment survival rate averaged a smaller 0.43 +/- 0.01 SD.

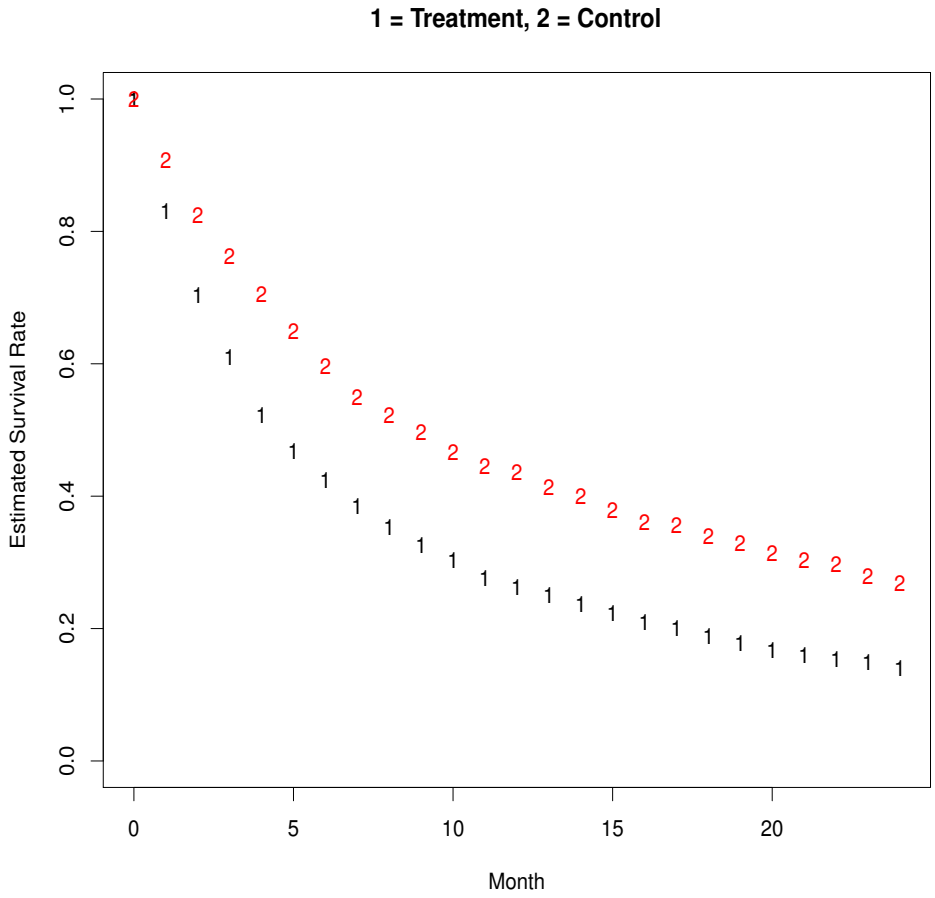

Figure 2: Estimated survival rates at every month from 0-2 years under treatment and control, using our method in the two simulated studies. We estimate the survival curves that would have resulted if all subjects in all studies had been assigned to treatment and then pooled, or if all subjects had been assigned control and then pooled. Unlike the pooling method previously discussed there is no longer a Simpson's paradox, because the curves correctly identify that treatment is harmful.

3.7 Correspondence to AKME

Xie and Liu (2005) introduced an Adjusted Kaplan-Meier Estimator (AKME) for comparing survival distributions between groups in nonrandomized or observational studies, which attempts to adjust for confounding by using covariates for propensity score modeling. The context is very different from synthesizing randomized clinical trials, but estimated survival curves can reduce to those derived here if each subject's covariate is a categorical variable identifying trial membership. In spite of this reduction we have seen that our counterfactual contrast approach to meta-analysis is quite nonstandard even in the binary outcome setting. For related methods of adjusting for confounders with matching or stratification we refer to Nieto and Coresh (1996), Amato (1998), Galimberti et al. (2002), or Winnett and Sasieni (2002).

3.8 Relaying Uncertainty

We have not presented confidence bands or pointwise confidence intervals for our method, because the standard approach in meta-analysis has been to instead plot Kaplan-Meier curves by treatment group (as in Figures 1 and 2) and also compute a p-value from a logrank test stratified by study (Lincoff et al., 2007; FDA, 2007). Our procedure suffices for allowing such a visual comparison of risk over time. If the follow-up time for a subject does not depend on treatment assignment (i.e., $z_i^T = z_i^C$), then analogous to our permutation approach with rosiglitazone in Section 2.5, we can form a null distribution for a test statistic such as a difference in estimated survival rates $F^T(s) - F^C(s)$ or a logrank statistic. By permuting treatment labels within each study and recomputing the statistic we can test the global null hypothesis of no treatment effect on failure times (i.e., $r_i^T = r_i^C$).

4 Summary

It is well-known that pooling subjects across randomized trials can bias conclusions, yet this procedure is often used when presenting Kaplan-Meier curves in systematic reviews, in part because there are few alternative methods for examining the time course of risk. To partially address this problem we have introduced a counterfactual model for time-to-event meta-analysis. Our method attempts to synthesize how risk changes over time for subjects assigned to treatment or control arms, and is immune to biases that can be

induced by unequal treatment assignment probabilities across trials. In spite of this robustness we generally only recommend combining trials in a time-to-event meta-analysis if they are relatively homogeneous and assess similar interventions in similar patient populations.

Appendix: Proof of Lemma 2

We prove the consistency result only for F^T, because the argument for F^C is identical. The proof is broken into seven steps. The heavy lifting is offloaded in Step 6 to a known result about the mapping we defined that takes (a^T, b^T) to f^T. Before beginning the proof we note that step functions a^T, b^T, a^C, and b^C are all nondecreasing, right continuous, and have left limits.

Step 1. We claim that (a^T, b^T) converges uniformly to (a^T_∞, b^T_∞) on $[0, \tau]$.

We prove the result for a^T, with the analysis for b^T_∞ being the same. The argument replicates several steps often used when proving the Glivenko-Cantelli theorem (Durrett, 2005).

Fix integer p. For $1 \leq j < p$, let $s_j = \inf\{s : a^T_\infty(s) \geq a^T_\infty(\tau)j/p\}$, and let $s_0 = 0$ and $s_p = \tau$. The pointwise convergence given in the lemma assumptions implies we can choose a point in the sequence of meta-analyses after which

$$|a^T(s_j) - a^T_\infty(s_j)| \leq 1/p \text{ and } |a^T(s_j-) - a^T_\infty(s_j-)| \leq 1/p$$

for $0 \leq j \leq p$. If $s \in (s_{j-1}, s_j)$ with $1 \leq j \leq p$, then using the monotonicity of a^T and a^T_∞, and the fact that $a^T_\infty(s_j-) - a^T_\infty(s_{j-1}) \leq a^T_\infty(\tau)/p \leq 1/p$, we have that

$$\begin{aligned} a^T(s) &\leq a^T(s_j-) \leq a^T_\infty(s_j-) + p^{-1} \leq a^T_\infty(s_{j-1}) + 2p^{-1} \leq a^T_\infty(s) + 2p^{-1}, \\ a^T(s) &\geq a^T(s_{j-1}) \geq a^T_\infty(s_{j-1}) - p^{-1} \geq a^T_\infty(s_j-) - 2p^{-1} \geq a^T_\infty(s) - 2p^{-1}. \end{aligned}$$

Therefore, we can choose a point in the sequence of meta-analyses after which $\sup_{s\in[0,\tau]}|a^T(s) - a^T_\infty(s)| \leq 2p^{-1}$. Because p is arbitrary, this gives the desired uniform convergence. □

DOI: 10.2202/1557-4679.1289

Step 2. We claim that (a^T_∞, b^T_∞) is right continuous on $[0, \tau]$.

We prove the result only for a^T_∞ as the argument for b^T_∞ does not change.

Fix time s and arbitrary $\epsilon > 0$. As a^T converges uniformly to a^T_∞ by Step 1 on $[0, \tau]$, we can choose a point in the sequence of meta-analyses such that $|a^T - a^T_\infty| \leq \epsilon/3$ on this interval. By the right continuity of distribution function a^T at this point in the meta-analysis sequence, we can choose δ_0 such that $0 < \delta < \delta_0$ implies $a^T(s+\delta) - a^T(s) \leq \epsilon/3$. Thus, for all $0 < \delta < \delta_0$ we have that

$$\begin{aligned}
&a^T_\infty(s+\delta) - a^T_\infty(s) \\
&= a^T_\infty(s+\delta) - a^T(s+\delta) + a^T(s+\delta) - a^T(s) + a^T(s) - a^T_\infty(s) \\
&\leq |a^T_\infty(s+\delta) - a^T(s+\delta)| + |a^T(s+\delta) - a^T(s)| + |a^T(s) - a^T_\infty(s)| \\
&\leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon,
\end{aligned}$$

which proves the desired right continuity. □

Step 3. We claim $(A^T(s), B^T(s))$ and $(A^T(s-), B^T(s-))$ converge pointwise in probability to $(a^T_\infty(s), b^T_\infty(s))$ and $(a^T_\infty(s-), b^T_\infty(s-))$ for $s \in [0, \tau]$.

We prove the result for A^T, with the argument for B^T being identical.

First, note that $A^T(s)$ is unbiased for $a^T(s)$, by the same reasoning as used in Lemma 1. By similar reasoning, $A^T(s-) = n^{-1} \sum_{i=1}^n (X_i/\pi_i) 1(Y_i^T < s)$ is unbiased for $a^T(s-) = n^{-1} \sum_{i=1}^n 1(y_i^T < s)$.

We next turn to the variance of $A^T(s)$. Note that $\mathrm{Cov}(X_i, X_j) \leq 0$ for $i \neq j$; if subjects i and j are in different studies then the covariance by assumption is zero, and if they are both in study l it is simple to compute as $\frac{m_l}{n_l}(\frac{m_l - 1}{n_l - 1} - \frac{m_l}{n_l}) < 0$. Thus,

$$\begin{aligned}
&\mathrm{Var}(A^T(s)) = \mathrm{Var} \quad n^{-1} \sum_{i=1}^n \frac{X_i}{\pi_i} 1(y_i^T \leq s) \Bigg) \\
&\leq n^{-2} \sum_{i=1}^n \frac{\mathrm{Var}(X_i)}{\pi_i^2} 1(y_i^T \leq s) = n^{-2} \sum_{i=1}^n \frac{\pi_i(1-\pi_i)}{\pi_i^2} 1(y_i^T \leq s) \\
&\leq n^{-2} \sum_{i=1}^n \frac{1}{\pi_i} = n^{-2} \sum_{l=1}^k n_l \frac{1}{\gamma_l} = n^{-1} \left(\sum_{l=1}^k \omega_l / \gamma_l \right).
\end{aligned}$$

By the lemma assumptions this tends to zero in our sequence of meta-analyses. An identical argument shows that $\mathrm{Var}(A^T(s-))$ tends to zero.

Therefore, $A^T(s) - a^T(s)$ and $A^T(s-) - a^T(s-)$ have mean zero and variance tending to zero, so Chebyshev's inequality implies that they converge to zero in probability. As $a^T(s)$ and $a^T(s-)$ converge to $a^T_\infty(s)$ and $a^T_\infty(s-)$ by the lemma assumptions, this implies the desired result of pointwise convergence in probability. □

Step 4. We claim that (A^T, B^T) converges uniformly in probability to (a^T_∞, b^T_∞) on $[0, \tau]$.

The argument in Step 1 shows that pointwise convergence on $[0, \tau]$ of bounded nondecreasing functions with left limits at s and $s-$ implies uniform convergence. Due to the pointwise convergence in Step 3 the argument can be repeated for A^T and B^T, only now with the uniform convergence result being in probability. □

Step 5. We set up probability spaces for random functions A^T, B^T, F^T.

Let Ω be a sample space for meta-analysis data $\{w_i, X_i, \Delta_i, Y_i\}_{i=1}^n$. We claim that in defining a probability model, we can use as a sigma field the collection 2^Ω of all subsets of Ω. Why? A probability measure P is indexed by the deterministic full counterfactual data $\{w_i, \delta_i^T, y_i^T, \delta_i^C, y_i^C\}_{i=1}^n$ and the joint distribution of treatment indicators $\{X_i\}_{i=1}^n$. Once this probability measure is fixed the observed data can only take finitely many values corresponding to the $\leq 2^n$ possibilities of treatment assignment. Hence, the probability measure reduces to one defined on a finite sample space, and for any $E \subset \Omega$ we have that $P(E)$ reduces to a finite sum over some of these $\leq 2^n$ elements $\omega \in \Omega$ with nonzero probability. Countable additivity likewise follows, meaning that $(\Omega, 2^\Omega, P)$ is a valid probability triple.

Now, let $D[0, \tau]$ denote the space of functions on $[0, \tau]$ that are right continuous with left limits. Define norm $\|g\| = \sup_{s \in [0,\ \tau]} |g(s)|$, metric $d(g_1, g_2) = \|g_1 - g_2\|$, and consider the Borel sigma-field containing all open sets.

We will view random functions A^T B^T, and F^T as random elements in $D[0, \tau]$. Considering a random function G as a map $G : \Omega \to D[0, \tau]$, the measurability of G easily follows, because for any $H \subset D[0, \tau]$ we have that $\{\omega : G(\omega) \in H\}$ is a subset of Ω, and hence is trivially in our sigma-field 2^Ω.

Thus, A^T B^T, and F^T are measurable when viewed as random functions taking values in $D[0, \tau]$, and probabilistic convergence in $D[0, \tau]$ is well-defined. For instance, because we endowed $D[0, \tau]$ with the supremum

norm, the uniform convergence of (A^T, B^T) in Step 4 can be expressed as $(A^T, B^T) \to_p (a^T_\infty, b^T_\infty)$ in $D[0,\tau]^2$. □

Step 6. We note (a^T_∞, b^T_∞) is in the domain of a continuous mapping ϕ.

Let $c^T(s) = 1 - a^T(s-)$, and consider the way in which we mapped the original functions (a^T, b^T) to the desired parameter f^T. This procedure can be viewed as a mapping $\phi : D[0,\tau]^2 \to D[0,\tau]$ defined by

$$\phi : \ (a,b) \to \lambda = \int \frac{db}{c} \to f = 1 - \prod(1 - d\lambda).$$

Section 3.9.4 of van der Vaart and Wellner (1996) shows that this mapping is defined and continuous (in fact Hadamard-differentiable) on a domain of type $\{(a,b) : c \geq \epsilon, \ \int |db| \leq M\}$ for given M and $\epsilon > 0$, at every point (a,b) such that function $1/c$ is of bounded variation. Restricting the functions to interval $[0,\tau]$, the lemma assumption that $a^T_\infty(\tau-) < 1$ gives that (a^T_∞, b^T_∞) is in this domain for $M \geq 1$ and sufficiently small ϵ. Also, as the sequence of meta-analyses increases, note that the convergence $(A^T, B^T) \to_p (a^T_\infty, b^T_\infty)$ in $D[0,\tau]^2$ implies (A^T, B^T) is in this domain with probability tending to one: as B^T is a subdistribution function we have $\int |dB^T| = \int dB^T \leq 1$ with probability one, and as function A^T is monotone we have $\inf_{s\in[0,\tau]}\{1 - A^T(s-)\} = 1 - A^T(\tau-) \to_p 1 - a^T_\infty(\tau-) > 0$. □

Step 7. We complete the proof using the continuous mapping theorem.

Because Step 4 shows that $(A^T, B^T) \to_p (a^T_\infty, b^T_\infty)$ in $D[0,\tau]^2$ and ϕ is a continuous mapping, it follows from the continuous mapping theorem that $F^T = \phi(A^T, B^T) \to_p \phi(a^T_\infty, b^T_\infty)$ in $D[0,\tau]$.

Because of the convergence $(a^T, b^T) \to (a^T_\infty, b^T_\infty)$ in $D[0,\tau]^2$ from Step 1, continuity of ϕ also implies $f^T = \phi(a^T, b^T) \to \phi(a^T_\infty, b^T_\infty)$ in $D[0,\tau]$.

The last two statements together give that $F^T - f^T \to_p 0$ in $D[0,\tau]$, implying the lemma. □

References

[1] Amato, A. (1988). Generalized Kaplan-Meier estimator for heterogeneous populations. *Communication in Statistics, Theory and Methods*, 17:263-286.

[2] Bracken, M. (2007). Rosiglitazone and cardiovascular risk. *New England Journal of Medicine*, 357:937-938.

[3] Chuang-Stein, C. and Beltangady, M. Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. *Pharmaceutical Statistics*, published online January 13, 2010.

[4] Cobitz, A., Zambanini, A., Sowell, M., Heise, M., Louridas, B., McMorn, S., Semigran, M., and Koch G. (2008). A retrospective evaluation of congestive heart failure and myocardial ischemia events in 14,237 patients with type 2 diabetes mellitus enrolled in 42 short-term, double-blind, randomized clinical studies with rosiglitazone. Pharmacoepidemiology and Drug Safety, 17:769-781.

[5] Higgins J.P.T., Green S. (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.2 [updated 2009]. The Cochrane Collaboration. Supplementary material on *Considerations and recommendations for figures in Cochrane reviews: graphs of statistical data.*

[6] Diamond, G.A., Bax, L., and Kaul, S. (2007). Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Annals of Internal Medicine*, 147:578-581.

[7] Ducharme, G.R. and Lepage, Y. (1986). Testing collapsibility in contingency tables. *Journal of the Royal Statistical Society Series B*, 48:197-205.

[8] Durrett, R. (2005). *Probability: Theory and Examples*. Brooks/Cole, Belmont, CA.

[9] FDA briefing document. www.fda.gov/ohrms/dockets/ac/07/briefing /2007-4308b1-02-fda-backgrounder.pdf

[10] Freedman, D.A. (2008). Randomization does not justify logistic regression. *Statistical Science*, 23:237-249.

[11] Galimberti S., Sasieni P., and Valsecchi M.G. (2002). A weighted Kaplan-Meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukemia. *Statistics in Medicine*, 21:3847-3864.

[12] Gill, R.D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18:1501-1555.

[13] Horvitz, D.G. and Thompson, D.J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663-685.

[14] Kaplan, E.L., and P. Meier. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association.* 53, 457-481.

[15] Lincoff, A.M., Wolski, K., Nicholls, S.J. and Nissen, S.E. (2007). Pioglitazone and risk of cardiovascular events in patients with type 2 diabetes mellitus: a meta-analysis of randomized trials. *Journal of the American Medical Association*, 298:1180-1188.

[16] Neyman, J. (1923) Sur les applications de la théorie des probabilitiés aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1-51.

[17] Nieto F.J. and Coresh J. (1996). Adjusting survival curves for confounders: A review and a new method. *American Journal of Epidemiology*, 143:1069-68.

[18] Nissen, S.E. and Wolski, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine,* 356:2457-2471.

[19] Psaty, B.M. and Furberg, C.D. (2007). Rosiglitazone and cardiovascular risk. *New England Journal of Medicine*, 356:24.

[20] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89:846-866.

[21] Rosenbaum, P.R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* 17:286304.

[22] Rücker G. and Schumacher M. (2008). Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology*, 8:34.

[23] Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, 13:238-241.

[24] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, NY.

[25] Winnett A and Sasieni P. (2002). Adjusted Nelson-Aalen Estimates with Retrospective Matching. *Journal the American Statistical Association*, 97:245-256.

[26] Xie J. and Liu C. (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine,* 20:3089-3110.