Combining cox regressions across a heterogeneous distributed research network facing small and zero counts

Martijn J. Schuemie\textsuperscript{1,2,3}, Yong Chen\textsuperscript{4}, David Madigan\textsuperscript{1,5}, and Marc A. Suchard\textsuperscript{1,3,6}

Abstract
Studies of the effects of medical interventions increasingly take place in distributed research settings using data from multiple clinical data sources including electronic health records and administrative claims. In such settings, privacy concerns typically prohibit sharing of individual patient data, and instead, cross-network analyses can only utilize summary statistics from the individual databases such as hazard ratios and standard errors. In the specific but very common context of the Cox proportional hazards model, we show that combining such per site summary statistics into a single network-wide estimate using standard meta-analysis methods leads to substantial bias when outcome counts are small. This bias derives primarily from the normal approximations of the per site likelihood that the methods utilized. Here we propose and evaluate methods that eschew normal approximations in favor of three more flexible approximations: a skew-normal, a one-dimensional grid, and a custom parametric function that mimics the behavior of the Cox likelihood function. In extensive simulation studies, we demonstrate how these approximations impact bias in the context of both fixed-effects and (Bayesian) random-effects models. We then apply these approaches to three real-world studies of the comparative safety of antidepressants, each using data from four observational health care databases.

Keywords
proportional hazards, meta-analysis, privacy preservation, Bayesian, distributed research networks

Introduction
Studies of the effects of medical intervention increasingly take place in distributed research settings using data from multiple clinical data sources. This is especially true in observational research, where studies draw on existing health care data such as electronic health records, administrative claims data, and registries. Such studies provide critical clinical knowledge, especially in settings where randomized trials prove impractical or overly costly. The recent emergence of distributed research networks, such as the Observational Health Data Sciences and Informatics (OHDSI),\textsuperscript{1} enables the use of data from hundreds of millions of patients across the world and can answer questions about relationships between exposures and outcomes, even for relatively rare exposures and outcomes.

\textsuperscript{1}Observational Health Data Sciences and Informatics, New York, NY USA
\textsuperscript{2}Janssen Research & Development, Titusville, NJ USA
\textsuperscript{3}Department of Biostatistics, University of California, Los Angeles, CA, USA
\textsuperscript{4}Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA
\textsuperscript{5}Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
\textsuperscript{6}Department of Human Genetics, University of California, Los Angeles, CA, USA

Corresponding author:
Martijn J. Schuemie, Janssen Research & Development, 1125 Trenton Harbourton Road, 08560, Titusville, NJ, USA.
Email: schuemie@ohdsi.org
Despite the promise of multi-site analyses, a number of analytic challenges commonly arise. First, sharing of individual patient data (IPD) rarely proves possible because of patient privacy concerns and local government regulations. Second, the nature of the data gathering process leads to censored observation periods that require a time-to-event analysis rather than simpler incidence rate estimation. Third, due to the observational nature of the data and the consequent potential for confounding, some correction for baseline differences between exposure groups always proves necessary, typically via stratifying, matching, or weighting by a propensity score or disease risk score. Fourth, since different data sites represent different patient populations, inter-site heterogeneity often arises. Finally, even though many of these health care databases contain the records of large numbers of patients, co-occurrences of even moderately rare exposures and/or outcomes often prove to be sparse to non-existent. For example, a recent study comparing the risk of angioedema following exposure to levetiracetam as compared to phenytoin used data from 10 databases covering over 300 million patients. The study identified 350,000 patient exposures to one or the other drug but only 125 patients experienced angioedema during exposure and several sites observed zero outcome events during exposure.

To avoid sharing of IPD, studies can at best receive only aggregate statistics from the study sites. Current practice typically estimates the hazard ratio (with standard error) at each site and then combines these estimates using a traditional random-effects meta-analysis. However, with small counts, the standard normal approximation of the per site likelihood function can break down and lead to substantial bias, especially when only one of the treatment arms yields zero outcome events. Wu et al. provide a recent, in-depth examination of the data condition under which the Cox partial likelihood function becomes monotonic. Recently, with the increasing research interest in rare outcomes in clinical applications such as drug safety, there is a growing body of work addressing the issue of combining small count data from multiple studies. Much of the literature eschews time-to-event data and concerns itself with sharing and combining 2×2 tables. Others have focused specifically on the analysis of time-to-event data but present methods that require sharing of IPD. Some authors have proposed the use of “risk sets,” sets of subjects having similar covariate values who are at risk at various time points. This approach primarily aims to account for heterogeneity in the selection of covariates, and in fact requires the normality assumption to hold even within risk sets.

In this paper, we aim to fill a methodological gap on evidence synthesis for time-to-event outcomes with small counts. Specifically, we propose and evaluate methods for combining evidence from Cox models across a distributed research network of databases without sharing IPD. In section “Likelihood approximations,” we propose and investigate four different approximations of the per site likelihood function: a normal approximation, a skew-normal approximation, a custom parametric approximation, and a grid approximation. We develop both a fixed-effect and a random-effects meta-analysis model using these approximations (section “Evidence synthesis via combining likelihoods from different sites”). For the latter, we build on the work by Seide et al. and utilize a Bayesian random-effects meta-analysis employing the different approximations of the likelihood. We evaluate these techniques using extensive simulations (section “Simulation studies”), and also demonstrate them in three case studies using four real-world databases (section “Applied examples”).

Methods

We assume a study is being executed across a network of databases. At each site, using IPD the shape of the log-likelihood function is approximated, and the parameters of the approximation are shared with the study coordinator. Centrally, a single estimate can then be synthesized from these per site approximation. For the rest of the paper, we focus on a single parameter Cox proportional hazards model, where the parameter quantifies the treatment effect comparing two exposures. We adjust for confounding variables by stratification or matching on propensity scores, which we compute using a large-scale propensity model using demographics as well as all prior conditions, exposures, procedures, measurements, etc. Such single parameter Cox proportional hazard models have proven to be an effective approach in many applications.

Likelihood approximations

Partial Cox likelihood

Suppose we have \( n \) subjects, with \( D \subset \{1, \ldots, n\} \) indicating those who experienced the outcome during their time at risk. \( z_1, \ldots, z_n \) is a binary variable indicating whether the subject was assigned the target or comparator treatment. The partial Cox likelihood can then be written as follows:

\[
L(\beta) = \prod_{i \in D} \frac{e^{\beta z_i}}{\sum_{k \in R(t_i)} e^{\beta z_k}}
\]

where \( R(t_i) \) is the risk set, the set of subjects still observed when subject \( i \) experienced the outcome, so those that did not
have the outcome or were censored before that time. When computing a stratified model, for example using propensity score strata, the risk set is further limited to only those subjects belonging to the same strata as $i$. The parameter to estimate is $\beta$, the treatment effect, which can be interpreted as the log of the hazard ratio.

In a distributed research network, we can compute the partial likelihood at each site and would like to share the shape of this function of parameter $\beta$, but without sharing any IPD. We therefore consider four ways to approximate and subsequently communicate this shape.

**Normal approximation**
A typical choice would be a simple quadratic approximation in the form of a normal function. Let $\phi(\beta)$ denote the standard normal probability density function

$$\phi(\beta, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-(\beta - \mu)^2/2\sigma^2}$$

We estimate $\hat{\mu}$ and $\hat{\sigma}$ using the mode and Fisher information of the Cox partial likelihood function. Note that if one or both of the treatment groups has zero outcomes, the mode is not defined and the normal approximation does not exist. Typically, analyses simply remove such sites altogether, potentially biasing the resulting meta-analytic estimate.

**Skew-normal approximation**
When the sample size is small, the Cox likelihood function can become skewed. To account for this, we evaluate the skew-normal function described by Azzalini, which generalizes the normal distribution by allowing for a degree of skewness. The skew-normal density function combines the cumulative normal distribution function:

$$\Phi(\beta, \mu, \sigma) = \int_{-\infty}^{\beta} \phi(t, \mu, \sigma)dt$$

with the normal density to yield:

$$f(\beta, \mu, \sigma, \alpha) = 2\phi(\beta, \mu, \sigma) \Phi (\alpha\beta, \mu, \sigma)$$

We estimate $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\alpha}$ using the procedure described below. Even in the presence of zero counts, the skew-normal can provide a reasonable approximation of the target likelihood function.

**Custom approximation**
Our investigations suggested that the skew-normal performed poorly in the face of severe skewness. We therefore propose a novel function, which we refer to as the “custom function”:

$$\ln(f(\beta, \mu, \sigma, \gamma)) = \ln(\left(1 + e^{-\gamma(\beta - \mu)}\right)^{\gamma})$$

We estimate $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\gamma}$ using the procedure described below. We note that the custom function defaults to a normal when skew is zero ($\hat{\gamma} = 0$). The motivation for this modification of the normal likelihood is to mimic the double exponential aspect which also features in the Cox likelihood function. Figure 1 shows this custom approximation function under various parameter choices. We have observed this family of functions perform well even in the presence of zero-counts.

**Grid**
Finally, we can communicate the (log) partial likelihood function by sampling values at predefined points in a one-dimensional grid of hazard ratios over a plausible range. Here we define the grid from a log hazard ratio as 1,000 equally spaced points spanning log (0.1) to log (10). Most if not all effect sizes of interest lie well within this range. We note that zero counts do not impact this approximation and increasing the grid size can provide an arbitrarily high-quality approximation.
Function fitting

We estimate the parameters for the skew-normal and custom function by minimizing the weighted sum of squares across $B$, a grid of values for $\beta$:

$$SS = \sum_{\beta \in B} w(\beta) \left\{ \log [L(\beta)] - \log [f(\beta)] \right\}^2$$

where $L(\beta)$ is the likelihood of $\beta$, $f(\beta)$ is our approximation function (skew-normal or custom), and $w(\beta)$ is the weight. The weight is simply the likelihood, truncated to a minimum value:

$$w(\beta) = \begin{cases} L(\beta) & \text{for } L(\beta) > 10^{-3}, \\ 10^{-3} & \text{otherwise} \end{cases}$$

We define $B$ as a predefined grid from log (0.1) to log (10), in 100 equal steps.

Evidence synthesis via combining likelihoods from different sites

Fixed-effect model

A fixed-effect model assumes the true hazard ratio $\mu$ is the same across all data sites. One way to estimate $\hat{\mu}$ is to optimize the overall likelihood, defined by the product of the per site likelihood:

$$L_{\text{all}}(\beta) = \prod_{n=1}^{N} L_n(\beta)$$

where $N$ is the number of data sites and $L_n(\beta)$ is the likelihood of log hazard ratio $\beta$ at site $n$. Here we will use the various approximations to convey $L_n(\beta)$ (normal, skew-normal, custom, and grid).

Once we have found the optimum, we can find the 95% confidence interval by finding $\beta$ with

$$\log [L_{\text{all}}(\hat{\mu})] - \log [L_{\text{all}}(\beta)] = \frac{q}{2}$$

where $q$ is the 95% quantile of the $\chi^2$ distribution with 1 degree of freedom.\textsuperscript{23}

Random-effect meta-analysis

A random-effects meta-analysis assumes the true hazard ratio $\theta_n$ from site $n$ is drawn from some distribution. Often, as we will do here, this distribution is assumed to be normal with mean $\mu$ and variance $\tau^2$:

$$\theta_n \sim N(\mu, \tau^2)$$

Because we expect to have little statistical power per database, we foresee problems in estimating $\tau$ accurately.\textsuperscript{24} We
therefore adopt the Bayesian approach of Seide et al., estimating $\hat{\mu}$ as the median of the posterior distribution of $\mu$. We identify the 95% credible interval as the 95% highest density interval, and compute a proxy of the standard error based on the credible interval and assuming a normal distribution. Similar to Seide et al., we use a half-normal prior with scale 0.5 on $\tau$. We assume a normal prior with a standard deviation of 2 on $\mu$, thus specifying 95% of the probability mass on the hazard ratio is between 0.02 and 50.40.

To learn the joint posterior distribution of $(\mu, \tau^2, \theta_1, \ldots, \theta_N)$, we employ Markov chain Monte Carlo (MCMC) via a random-scan Metropolis-within-Gibbs sampling scheme. The scheme interleaves Gibbs sampling from full conditional distributions $p(\mu | \tau^2, \theta_1, \ldots, \theta_N)$ and $p(\tau^2 | \mu, \theta_1, \ldots, \theta_N)$ with separate random-walk Metropolis–Hastings transition kernels on $\theta_n$ for all $n$. These latter kernels have auto-tuning scale constants to efficiently sample from a variety of posteriors. To approximate each posterior, we simulate MCMC chains for 1.1 million steps, discarding the first 0.1 million steps as burn-in and sub-sample every 100 steps to decrease auto-correlation between samples. For our applications, these settings generate effective sample sizes of well over 1,000 across all model parameters. We make our implementation available as an R package EvidenceSynthesis that relies on the popular Bayesian sampling software BEAST.

Simulation studies

Simulation settings and measures to compare performance of different methods

We perform two separate simulation studies, one assuming fixed-effects only, and one where $\tau$ is allowed to be >0. The reason for having two separate simulation studies is the computational expense of our random-effects approach, meaning we had to be conservative in the number of simulations. The simulation framework is described in Appendix A in the online supplemental material. We consider the following simulation parameters:

- **Treated fraction**: The fraction of the study population treated with the target treatment. The remainder of the population is assumed to have comparator exposure.
- **Hazard ratio**: The true (mean) hazard ratio.
- **N sites**: The number of database sites.
- **Max n**: The maximum number of subjects per site (size is sampled from unif(1000, Max n)).
- **N strata**: The number of propensity or disease risk score strata. The baseline risk is simulated to be constant within strata but differs across strata.
- **$\tau$**: The true standard deviation of the random-effects distribution.

For each parameter, we select several values and then create the full factorial combination of all values. Each unique combination of parameter values forms a "simulation scenario," and is repeated 1,000 times, allowing us to compute the following metrics:

- **Coverage**: The fraction of times the true hazard ratio is within the 95% confidence or credible interval.
- **Bias**: The mean of the difference between the log of the estimated hazard ratio and the log of the true hazard ratio.
- **MSE**: The mean of the square of the difference between the log of the estimated hazard ratio and the log of the true hazard ratio.
- **Precision**: The geometric mean of the precision ($1/(\text{standard error})^2$).
- **Non-estimable**: The fraction of simulation iterations where an estimate could not be produced, for example, because all sites had zero counts.

Simulation results

In total, we evaluated 540 and 216 unique simulation parameter combinations in the fixed-effects and random-effects simulation studies, respectively, and for each combination, we computed the performance metrics. Our online interactive visualization at https://data.ohdsi.org/NonNormalEvidenceSynthesisSimulations/ enables a complete review of the results.

Fixed-effects simulations

The violin plots in Figure 2 show the distributions of performance metrics across the fixed-effects simulations. Of the five simulation parameters that we varied, the treatment fraction proved to be most important in explaining differences in performance. That is why we show this parameter on the horizontal axis. When the treatment groups are of unequal size
(i.e. the treatment fraction moves away from 0.5), the normal approximation shows increasing bias and lower coverage of the 95% confidence interval. All other approximations appear essentially unbiased.

**Random-effects simulations**

Although in the random-effects meta-analyses the treated fraction was still an important determinant of performance, we chose to highlight the number of sites in Figure 3. This demonstrates that the low coverage for the normal approximation decreases further as the number of sites increases. One can also see that the non-normal approximations remain unbiased at all simulation parameter values.
Custom versus grid approximation
We conclude that the custom and grid approximations both represent good candidates to minimize bias. Table 1 explores how these two approaches compare, showing the means and mean differences for the various metrics. We see the grid approximation achieves slightly lower bias and MSE in all simulations, and slightly higher precision when using a fixed-effects model. Coverage is largely comparable.

Figure 3. Distributions of performance metrics of random-effects meta-analyses using the various likelihood approximations across the 216 distinct random-effects simulation scenarios. The distributions are stratified by the number of sites, which is one of the simulation parameters that was varied. Dashed lines indicate reference values of a perfect performance, where applicable. All meta-analyses were Bayesian, except the “traditional” meta-analyses that used the traditional frequentists random-effects meta-analysis, using a normal approximation. MSE: mean squared error.
Study design and data

Although regulators require all medical treatments to undergo extensive evaluation in clinical trials prior to marketing, ensuring safety post-marketing remains a significant societal priority; rare adverse effects may fail to appear in typically sized clinical trials but result in a significant post-marketing public health challenge. Large-scale, routinely collected health care data can shed light on previously unknown side effects, including rare outcomes. For example, in a prior large-scale study,28 we used four US administrative claims databases to estimate the effect of depression treatments, comparing 17 treatments for 22 outcomes of interest. From this study, we picked three examples to reflect various situations where we encountered small and zero counts:

- **Example 1:** Amitriptyline (target) versus citalopram (comparator) for the risk of acute liver injury. We selected this scenario as it had non-zero counts across all four databases, and the highest mean skew (here interpreted at the ratio between the upper part and the lower part of the 95% confidence interval).
- **Example 2:** Nortriptyline (target) versus duloxetine (comparator) for the risk of acute liver injury. We selected this scenario as it had two databases with zero counts and the highest average skew.
- **Example 3:** Nortriptyline (target) versus venlafaxine (comparator) for the risk of decreased libido. We selected this scenario as it had two databases with zero counts in the target and the highest counts in the comparator.

In all three examples, we compared new users of each drug, requiring 365 days of continuous observation prior to treatment initiation as well as a prior diagnose of major depression, but excluding people who had the outcome before treatment initiation. We define time-at-risk to start on the day of treatment initiation and stop when treatment stopped, allowing for a 30-day gap in treatment continuation. We identified exposures as any dispensing of a drug containing the mentioned ingredient. We defined the acute liver injury outcome as a diagnosis of acute liver injury in an emergency room or inpatient setting.29 We defined decreased libido as any occurrence of a diagnosis indicating decreased libido. To address confounding, we utilized large-scale propensity models using demographics as well as all prior conditions, exposures, procedures, measurements, etc., and utilized \( L_1 \) regularization for the estimation.18 We then used these propensity scores to stratify the target and comparator cohorts in 10 equally sized strata and conditioned the proportional hazards outcome models on those strata.

We used four observational health care databases:

- IBM MarketScan® Commercial Claims and Encounters (CCAE) is an administrative health claims database for active employees, early retirees, Consolidated Omnibus Budget Reconciliation Act continues, and their dependents insured by employer-sponsored plans (individuals in plans or product lines with fee-for-service plans and fully capitated or partially capitated plans). At the time of analysis, CCAE contained 131 million patients.
- IBM MarketScan® Medicare Supplemental Beneficiaries (MDCR) is an administrative health claims database for Medicare-eligible active and retired employees and their Medicare-eligible dependents from employer-sponsored

---

**Table 1.** Comparison of performance metrics between the custom and grid approximations in the fixed-effects simulations using a fixed-effects model, and in the random-effects simulations using a random-effects model. The \( p \)-value corresponds to a two-sided \( t \)-test for paired samples.

| Type               | Metric  | Mean (custom) | Mean (grid) | Mean diff. | \( p \)  |
|--------------------|---------|---------------|-------------|------------|--------|
| Fixed-effects      | Bias    | 0.013         | 0.000       | 0.013      | <0.001 |
| Fixed-effects      | Coverage| 0.951         | 0.952       | −0.000     | 0.085  |
| Fixed-effects      | MSE     | 0.101         | 0.096       | 0.005      | <0.001 |
| Fixed-effects      | Non-estimable | 0.008   | 0.015       | −0.007     | <0.001 |
| Fixed-effects      | Precision| 20.725        | 21.640      | −0.915     | <0.001 |
| Random-effects     | Bias    | 0.005         | 0.000       | 0.004      | <0.001 |
| Random-effects     | Coverage| 0.946         | 0.948       | −0.002     | <0.001 |
| Random-effects     | MSE     | 0.190         | 0.178       | 0.012      | <0.001 |
| Random-effects     | Non-estimable | 0.000   | 0.000       | 0.000      |        |
| Random-effects     | Precision| 11.125         | 11.126      | −0.001     | 0.975  |

\( MSE \): mean square error.
supplemental plans (predominantly fee-for-service plans). Only plans where both the Medicare-paid amounts and the employer-paid amounts were available and evident on the claims were selected for this database. At the time of analysis, MDCR contained 9.6 million patients.

- IBM MarketScan® Multi-state Medicaid (MDCD) is an administrative health claims database for the pooled health care experience of Medicaid enrollees from multiple states. At the time of analysis, MDCD contained 21.6 million patients.

- Optum® De-Identified Clininformatics® Data Mart Database (Optum) is an adjudicated administrative health claims database for members with private health insurance, who are fully insured in commercial plans or in administrative services only, Legacy Medicare Choice Lives (prior to January 2006), and Medicare Advantage. At the time of analysis, Optum contained 74.7 million patients.

Schuemie et al.28 provide full details of this study including the protocol.

Because we have direct access to all four databases, we were able to pool the IPD extracted from these databases in a single regression model to provide a benchmark of the analysis based on pooled data. In an ordinary distributed network setting, this would not be possible.

Results

The table insets in Figure 4 show the number of subjects and outcome events for all three examples in the four databases in the two exposure cohorts. Although the number of included subjects is often large, the number of events observed during exposure is low, and sometimes zero. Figure 4 furthermore shows that in many scenarios the normal approximation provides a poor fit for the actual likelihood functions. In both Examples 2 and 3 in the MDCD and MDCR databases, there is no well-defined maximum likelihood estimate, and no normal approximation exists. As a consequence, we removed these databases from the meta-analysis for these examples. In contrast, both the skew-normal and custom approximations hew closely to the true likelihood. Because we designed the custom approximation to more easily approximate functions with extreme skew, the fitted parameters were smaller than those of the skew-normal. For example, for Example 2 in the MDCD database, the fitted parameters for the custom approximation were $\mu = -3.72$, $\hat{\sigma} = 8.39$, and $\gamma = 0.47$, while for skew-normal the fitted parameters were $\mu = 0.27$, $\hat{\sigma} = 477.64$, and $\gamma = -474.45$. However, these larger values did not appear to affect the accuracy of the approximation. Since the grid approximation follows the true likelihood in the plotted range, we do not show it here.

Table 2 shows the hazard ratio estimates generated using the various meta-analysis algorithms. The gold standard, denoted in bold font, is a Cox regression of the pooled data, which in a real application would not be available because patient-level data cannot be shared. For Example 1, both the traditional fixed-effects analysis as well as the traditional (non-Bayesian) random-effects meta-analysis,5 both using normal approximations, produce confidence intervals that do not include 1. However, as the gold standard demonstrates, this statistical significance derives solely and artifactually from the normal approximation. The normal approximation in a Bayesian random-effects meta-analysis does produce a confidence interval that includes 1, but shifts relative to the gold standard. In contrast, the skew-normal, custom, and grid approximations all produce meta-analytic estimates close to the gold standard. For Examples 2 and 3, the normal

| Method                     | Example 1   | Example 2   | Example 3   |
|----------------------------|-------------|-------------|-------------|
| Pooled fixed-effects       | 1.59 (0.91–2.70) | 0.84 (0.28–2.06) | 0.76 (0.55–1.03) |
| Traditional fixed-effects  | 1.77 (1.01–3.11) | 0.81 (0.25–2.60) | 0.83 (0.61–1.14) |
| Skew-normal fixed-effects  | 1.59 (0.88–2.72) | 0.83 (0.28–2.01) | 0.77 (0.55–1.03) |
| Custom fixed-effects       | 1.60 (0.90–2.72) | 0.83 (0.28–2.05) | 0.76 (0.55–1.03) |
| Grid fixed-effects         | 1.59 (0.91–2.69) | 0.84 (0.28–2.05) | 0.76 (0.56–1.02) |
| Pooled random-effects      | 1.45 (0.69–3.00) | 0.78 (0.26–2.28) | 0.63 (0.28–1.19) |
| Traditional random-effects | 1.77 (1.01–3.11) | 0.81 (0.25–2.60) | 0.83 (0.61–1.14) |
| Normal random-effects      | 1.64 (0.78–3.43) | 0.81 (0.23–2.83) | 0.83 (0.44–1.57) |
| Skew-normal random-effects | 1.43 (0.66–2.96) | 0.77 (0.24–2.26) | 0.64 (0.29–1.20) |
| Custom random-effects      | 1.44 (0.67–2.92) | 0.78 (0.24–2.36) | 0.64 (0.28–1.18) |
| Grid random-effects        | 1.43 (0.67–2.94) | 0.79 (0.24–2.39) | 0.63 (0.28–1.18) |

"Traditional" analyses are non-Bayesian and use normal approximations.
approximations produce wider confidence intervals than the gold standard and the other approximations, probably because of the two excluded databases.

These examples illustrate the profound clinical implications involved with these analytic choices. For our first example, the traditional approach incorrectly yields a statistically significant effect, potentially causing unwarranted concerns over the safety of amitriptyline. For the other two examples, more uncertainty would have remained concerning the potential

Figure 4. Three real world examples of evidence synthesis across a network of four databases. For each database, the plot shows the log-likelihood function of the Cox regression, as well as the various approximations. All log-likelihood values were normalized so the maximum within the plotted range equals 0. Zero counts were present for both Examples 2 and 3 in the MDCD and MDCR database, where no normal approximation was made. The grid approximation by definition follows the true likelihood in the plotted range and is therefore not shown.
magnitude of the effect size. Without sharing IPD, the various non-normal approximations were able to produce estimates in line with a gold standard that would not be available in most real-world studies.

Discussion
Our simulations show that the use of traditional meta-analytic techniques in distributed research networks can lead to substantial bias and low confidence interval coverage, especially when the compared treatment groups differ in size, and when the number of sites is large ($n > 10$). The clinical applications we presented confirm that this issue exists in real clinical studies, albeit in a less extreme than what the simulation studies suggest. Since our clinical applications necessarily included relatively small numbers of sites, we draw little comfort from this attenuation; clinical studies of the future will draw ever-increasing numbers of sites, and standard meta-analytic approximations have great potential to provide misleading results.

We found that combining appropriate approximations of the per database partial likelihood can effectively eliminate the bias from traditional meta-analyses facing small or zero counts. Such a strategy avoids the sharing of IPD, and hence is suitable for multi-site collaborations, especially for settings such as OHDSI. The skew-normal, custom, and grid approximations show good performance, in both fixed- and random-effects settings. We believe all three could be used in practice, although the custom approximation may provide the most compact representation to be communicated between sites. The normal approximation, as suspected, performed poorly in various simulated and real scenarios.

The non-normal approximations evaluated here can account for two situations where the normal approximation fails. First, when counts are low (but not zero), the likelihood will have a maximum but may be skewed, and second, when counts are zero and no maximum exists, but information about the parameter of interest still exists.

Even though our approach does not assume the per site likelihood is normally distributed, the Bayesian meta-analysis does model the between-site random effect as arising from a normal distribution. The true distribution of the random effect is unknowable, but prior research suggests that the choice of distribution family may not be that important.\textsuperscript{30-34}

Our approximations capture the profile likelihood for the parameter of interest, typically the effect on the outcome. The regression model itself can be as complex as it needs to be, in our case, for example, using stratification on a propensity score, or including additional covariates that are not of interest, but adjust for confounding. Note that when computing the profile likelihood of one parameter in a multi-parameter model the “nuisance” parameters should be maximized for all sampled values of the parameter of interest. We currently specifically focus on Cox regression. However, this same procedure may also be useful for other types of models, including Poisson and logistic regression.

In this paper, we considered hazard ratios in the 0.1–10 range as we believe this is realistic. However, the model can easily extend to different ranges as necessary.

As the sample size becomes smaller, the prior will play a more important role in defining the posterior. In Bayesian meta-analysis in general, as well as our approach here, sample size can be defined as the sample size per site, as well as the size of the sample of sites. The default priors for $\mu$ and $\tau^2$ we evaluated (see the “Methods” section) are fairly uninformative, leading to wider credible intervals of our meta-analytic estimates. An exploration of the effect of choice of parameter distributions (Appendix A, section 5 in the online supplemental material) shows that especially when the number of sites is small ($n < 5$), using a more informed prior for $\tau^2$ can reduce the width of the credible intervals, as well as move the posterior median. Whether such a stronger prior is justified depends on external knowledge of the heterogeneity of the data network.

Our current work focuses on the one-dimensional likelihood function. After proper propensity score matching or stratification, this approach has utility in many pharmacoepidemiological settings. Future research may explore higher-dimensional approximations, although this may prove intractable with increasing dimensionality. An alternative direction could be to extend our prior work fitting multi-variable time-to-event models in distributed settings, which restricts to a neighborhood of a specific point on the likelihood curve.\textsuperscript{6}

Although our approach was designed specifically for situations where the sample size per site is small, it is unclear at what sample size it no longer is necessary. Our simulation results stratified by sample size (Appendix A sections 4.4 and 4.9 in the online supplemental material) suggest that even for larger sample sizes the bias when using normal approximations can still be considerable. We therefore believe using a non-normal approximation is always advisable when performing time-to-event analyses in distributed research settings. To support this practice and perform the requisite calculations, we have created the EvidenceSynthesis R package, which is freely available on the Comprehensive R Archive Network.

Declaration of conflicting interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is partially supported through National Institutes of Health grant U19AI135995 and Food and Drug Administration contract “OHDSI-based FDA BEST Community Engagement and Development Coordination Center.”

ORCID iDs
Martijn J. Schuemie https://orcid.org/0000-0002-0817-5361
Marc A. Suchard https://orcid.org/0000-0001-9818-479X

Supplemental material
Supplemental material for this article is available online.

References
1. Hripcsak G, Ryan PB, Duke JD et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci 2016; 113: 7329–7336.
2. Rosenbaum P and Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41–55.
3. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol 1976; 104: 609–620.
4. Duke JD, Ryan PB, Suchard MA et al. Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network. Epilepsia 2017; 58: e101–e106.
5. DerSimonian R and Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7: 177–188.
6. Duan R, Luo C, Schuemie MJ et al. Learning from local to global—an efficient distributed algorithm for modeling time-to-event data. J Am Med Inform Assoc 2020; 27(7): 1028–1036. https://doi.org/10.1093/jamia/ocaa044
7. Wu J, de Castro M, Schifano ED et al. Assessing covariate effects using jeffreys-type prior in the Cox model in the presence of a monotone partial likelihood. J Stat Theory Pract 2018; 12: 23–41.
8. Sweeting MJ, Sutton AJ and Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004; 23: 1351–1375.
9. Friedrich JO, Adhikari NK and Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. BMC Med Res Methodol 2007; 7: 1–6.
10. Tian L, Cai T, Pfeffer MA et al. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2×2 tables with all available data but without artificial continuity correction. Biostatistics 2009; 10: 275–281.
11. Liu D, Liu RY and Xie Mg. Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events. J Am Stat Assoc 2014; 109: 1450–1465.
12. Cafri G, Banerjee S, Sednakyan A et al. Meta-analysis of survival curve data using distributed health data networks: Application to hip arthroplasty studies of the international consortium of orthopaedic registries. Res Synth Methods 2015; 6: 347–356.
13. Gronsbell J, Hong C, Nie L et al. Exact inference for the random-effect model for meta-analyses with rare events. J Stat Med 2020; 39: 252–264.
14. Siannis F, Barrett J, Farewell V et al. One-stage parametric meta-analysis of time-to-event outcomes. Stat Med 2010; 29: 3030–3045.
15. Yuan X and Anderson SJ. Meta-analysis methodology for combining treatment effects from Cox proportional hazard models with different covariate adjustments. Biom J 2010; 52: 519–537.
16. Li X, Fireman BH, Curtis JR et al. Validity of privacy-protecting analytical methods that use only aggregate-level information to conduct multivariable-adjusted analysis in distributed data networks. Am J Epidemiol 2019; 188: 709–723.
17. Seide SE, Röver C and Friede T. Likelihood-based random-effects meta-analysis with few studies: Empirical and simulation studies. BMC Med Res Methodol 2019; 19: 16.
18. Tian Y, Schuemie MJ and Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol 2018; 47: 2005–2014.
19. Suchard MA, Schuemie MJ, Krumholz HM et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis. Lancet 2019; 394: 1816–1826.
20. Hripcsak G, Suchard MA, Shea S et al. Comparison of cardiovascular and safety outcomes of chlorthalidone vs hydrochlorothiazide to treat hypertension. JAMA Intern Med 2020; 180: 542–551.
21. You SC, Rho Y, Bikdeli B et al. Association of ticagrelor vs clopidogrel with net adverse clinical events in patients with acute coronary syndrome undergoing percutaneous coronary intervention. JAMA 2020; 324: 1640–1650.
22. Azzalini A. The skew-normal and related families. Institute of Mathematical Statistics Monographs. Cambridge, UK: Cambridge University Press, 2013. doi:10.1017/CBO9781139248891.
23. Venzon D and Moolgavkar S. A method for computing profile-likelihood-based confidence intervals. J R Stat Soc C: Appl Stat 1988; 37: 87–94.
24. Friede T, Rover C, Wandel S et al. Meta-analysis of few small studies in orphan diseases. Res Synth Methods 2017; 8: 79–91.
25. Kruschke JK. Doing Bayesian data analysis: A tutorial with R and BUGS. Burlington, MA: Academic Press, 2011. ISBN 9780123814852 0123814855. http://www.amazon.com/Doing-Bayesian-Data-Analysis-Tutorial/dp/0123814855.

26. Liu JS, Wong WH and Kong A. Covariance structure and convergence rate of the Gibbs sampler with various scans. J R Stat Soc B: Methodol 1995; 57: 157–169.

27. Suchard MA, Lemey P, Baele G et al. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. Virus Evol 2018; 4: vey016.

28. Schuemie MJ, Ryan PB, Hripcsak G et al. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philos Trans A Math Phys Eng Sci 2018; 376. https://doi.org/10.1098/rsta.2017.0356

29. Udo R, Maitland-van der Zee AH, Egberts TC et al. Validity of diagnostic codes and laboratory measurements to identify patients with idiopathic acute liver injury in a hospital database Pharmacoepidemiol Drug Saf 2016; 25: 21–28.

30. McCulloch CE and Neuhaus JM. Prediction of random effects in linear and generalized linear models under model misspecification. Biometrics 2011; 67: 270–279.

31. Neuhaus JM, McCulloch CE and Boylan R. A note on type II error under random effects misspecification in generalized linear mixed models. Biometrics 2011; 67: 654–656.

32. McCulloch CE and Neuhaus JM. Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. Stat Sci 2011; 26: 388–402. https://doi.org/10.1214/11-STS361.

33. Litière S, Alonso A and Molenberghs G. Type I and type II error under random-effects misspecification in generalized linear mixed models. Biometrics 2007; 63: 1038–1044.

34. Neuhaus JM, McCulloch CE and Boylan R. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. Stat Med 2013; 32: 2419–2429.