SENTINEL METHODS REPORT

SAFETY SIGNALING METHODS FOR SURVIVAL OUTCOMES TO CONTROL FOR CONFOUNDING IN THE MINI-SENTINEL DISTRIBUTED DATABASE

Prepared by: Andrea J Cook, PhD[1,2], Robert D Wellman, MS[1], Rima Izem, PhD[3], Rongmei Zhang, PhD[3], Michael Nguyen, MD[4], Ram C Tiwari, PhD[3], Susan R. Heckbert, MD, PhD[5], Susan Gruber, PhD[6], and Jennifer C Nelson, PhD[1,2]

Author Affiliations: 1. Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, Seattle, WA 2. Department of Biostatistics, University of Washington, Seattle, WA 3. Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 4. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 5. Department of Epidemiology, University of Washington, Seattle, WA 6. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA

August 14, 2018

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I. This project was funded by the FDA through HHS Mini-Sentinel contract number HHSF223200910006I.

Sentinel Methods Report

Safety Signaling Methods for Survival Outcomes to Control for Confounding in the Mini-Sentinel Distributed Database

Table of Contents

Table of Contents: Tables and Figures

I. BACKGROUND, OBJECTIVES AND RECOMMENDATIONS FROM THE TASK ORDER

This is the final report summarizing the goals, methodologies and findings of the Sentinel Survival Workgroup 2 task order. The main objective of this working group was to evaluate statistical methods for controlling confounding when using Sentinel claims data to estimate the hazard ratio of a chronic exposure on a binary, potentially rare, outcome. All analyses use time to event Cox Proportional Hazards models that can account for varying follow-up time. Different methods to control for confounding are considered in this work order, more specifically propensity score regression and propensity score stratification. These methods are considered in a distributed data setting with a constraint of no sharing of subject level information between sites but where sharing of some summary level information is possible. The test cases, interim and final goals of this working group are shown in **Figure 1** and discussed in the remainder of this section.

This report compares the performance of multiple methods with simulated data in realistic scenarios. To generate these realistic scenarios, this workgroup built a simulation tool described in **Section II**. The simulation framework is a two-stage process. The first stage extracts summary level information from subject-level real world data. The second stage generates subject level data from summary level information. The resulting simulated data mimics real world settings like Sentinel with complex relationships between confounders, exposure of interest, and outcomes. Key aspects include simulating numerous confounders that can be related to each other, different exposure and confounder relationships, and outcome relationships with complex censoring and outcome distributions. This simulation framework was used to generate data to compare different methods throughout the rest of the report.

In **Section III** we present different methods for control of confounding and estimating risk for the non-distributed data setting. Then, we report performance of these methods in **Section IV** from a simulation evaluation. The simulation specifications were anchored to two real examples: 1) ACE and angioedema and 2) Rivaroxaban and Ischemic Stroke, but only for non-distributed data setting methods. In **Section V** we present methods viable for the distributed data setting extending the most promising non-distributed data approaches. Finally, **Section IV** shows a simulation evaluation using the real data examples as anchors.

Appendix C shows results of related work on binary outcomes rather than time to event outcome. This work evaluates statistical properties of the proposed propensity score methods in this task order as well as exposure matching and IPTW methods that are not included in this task order.

Some of the methods that are compared are not fully implemented in Sentinel. Thus, the regulatory aim of this workgroup is to provide advice on which methods are appropriate for certain specifications of exposure, outcome prevalence and variability between sites. Further, the new simulation tool can now be replicated by researchers in both Sentinel and outside of Sentinel who are interested in comparing methods for survival outcomes. By allowing for correlated and complex confounder relationships the tool makes it easier for researchers to simulate more realistic data including claims data.

We focused on methods that estimate a conditional hazard ratio (HR) and not methods that estimate a marginal time average HR such as exposure matching or inverse probability of treatment weighting

(IPTW). Methods which estimate marginal HR are investigated in other Sentinel working groups[1]. Simulations in **Section IV** and **Section VI** support using 10 PS strata rather than 5 PS strata as a preferred default for PS stratification. Similarly, in PS regression, using PS indicators for 10 strata performed well and using splines on the PS score had better properties.

To determine the scope of methods considered in this workgroup, we relied on what was known in the literature and filled some gaps. Austin 2014(41) compared exposure matching, IPTW, stratification, and regression on the propensity score (PS) for estimating marginal and conditional HRs. The simulations in the paper showed that IPTW and exposure matching were unbiased for the marginal time average HR, with IPTW being more efficient than 1:1 PS exposure matching. Moreover, the author showed that stratification and regression were biased in estimating the conditional HR.

A limitation of the findings in this paper is that the author investigated PS stratification with only 5 strata which probably explains the large observed biases due to residual confounding. Using PS quintiles or 5 PS strata is a recommendation from the past 20 years because it can eliminate up to 90% of bias due to measured confounding. However, recent work has shown that 5 strata is often not sufficient to control for confounding (33) with recommendation for using as many as 10 strata in an analysis.(46) Larger number of strata can reduce up to 95% of the bias, a more current standard nowadays with richer datasets available to estimate smaller safety risks considered the standard.(8, 36)(4) Our simulation evaluation in **Sections IV** and **VI** show that we have similar results reducing bias between 96-99% when stratifying on 10 PS strata by site. We recommend from both literature and our simulation evaluation presented in this report that at least 10 PS strata should be used and sensitivity analyses in which 15 or 20 PS strata are used to assess if residual confounding still persists. However, some caution should be taken for too many strata if most strata become too sparse.(39)

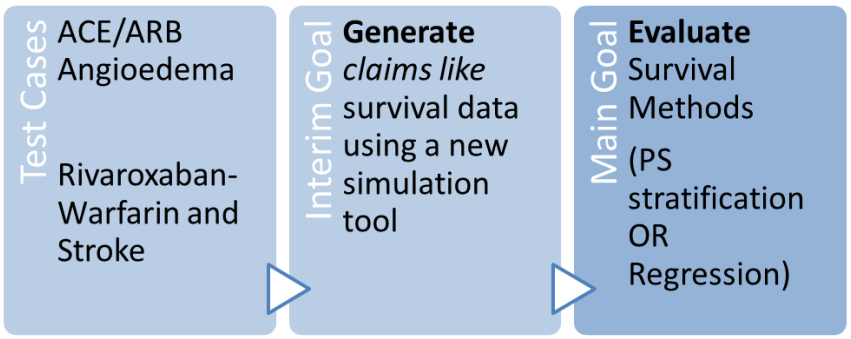
Another issue with methods in Austin was the application of PS regression adjustment using either a linear PS term or adjusting for 5 propensity score stratum indicators. Assuming a linear relationship between the propensity score and the log hazard is likely not the correct model specification and therefore may lead to residual confounding. We show throughout this document that adjusting for splines was a straightforward approach to fix this issue. However, in the distributed data setting splines are likely not feasible since they require subject level propensity score data. Adjusting for 5 propensity score indicators may not be enough to provide the flexibility needed in the propensity score model, depending upon the amount of confounding and distribution of the propensity score. This mirrors the issues with bias when only stratifying by 5 propensity score strata. We find in our simulation study that at least 10 strata were needed when doing propensity score adjustment. Therefore, similar to propensity score stratification, when doing propensity score adjustment, we recommend use of a flexible modeling approach like splines, or to use at least 10 quantile strata for adjustment, with sensitivity analyses for 15 or 20 strata.

A final issue with Austin[1] is that his simulation was not anchored in real examples and did not consider censoring. This motivated our analyses to mimic extensive censorings observed in post-market safety surveillance in electronic healthcare data.

[1] https://www.sentinelinitiative.org/sentinel/methods/evaluation-propensity-score-based-methods-sentinel-study-settings-using-simulation

This work follows from previous a Mini-Sentinel workgroup (Survival Workgroup I; Task Order PI: Cook(44)), further described in the **Appendix A**.

Figure 1. Test cases, interim and main goals of the survival working group



II. SIMULATING REALISTIC TIME-TO-EVENT DATA USING SHAREABLE SUMMARY INFORMATION THAT PROTECTS DATA PRIVACY

A. INTRODUCTION

There has been a rise in the use of large complex observational cohorts to address comparative effectiveness and safety research questions, primarily due to the development of collaborative research and data networks. Examples of such national research networks include the Health Care Systems Research Network (HCSRN), the Food and Drug Administration's Sentinel Initiative (FDA's Sentinel Initiative), and the emergent Patient-Centered Outcomes Research Network (PCORnet). Each of these data networks is comprised of an assemblage of partner organization that collect electronic health and claims data as part of their operations, but not necessarily for the purpose of research or medical product regulation. The FDA's Sentinel initiative, with its focus on postmarket safety surveillance, is especially interesting from a statistical perspective.

The standard approach to evaluating statistical methods is to conduct simulation experiments probing for scenarios that result in loss of accuracy, precision and/or power. As discussed by Franklin, et al. (2014)(42) there are a number of reasons why it may be advantageous to connect the data generation mechanism used in a simulation to an actual empirical study. Using data from existing empirical studies to inform data generation is useful in narrowing the focus of a simulation to those issues, both known and unknown, that are most salient to the particular context in which studies are being conducted. For postmarket safety surveillance efforts taking place within Sentinel, some important context-specific issues that may warrant consideration include: Rare events, rare exposures, confounding of varying degree and by a potentially large number of variables, complex relationships between confounding variables, covariate dependent censoring times and idiosyncratic prescribing patterns across institutions.

A further consideration when conducting research in a distributed data network is the need to maintain the privacy of patient-level data and to conduct research in such a way that sharing of data, which may be considered proprietary, is minimized. Franklin et al (2014)(42) presented plasmode simulation with patient-level data from empirical cohort studies as an effective way to mimic the complexity of the observed data for simulation evaluations. However, when privacy or proprietary concerns preclude

sharing subject-level data, plasmode simulation may not be feasible. In this paper we propose techniques for simulating data that reflect important and unique aspects of the data from an empirical surveillance study. Summary information ranging from basic descriptive statistics to model coefficients can be estimated through distributed queries to any number of data partner sites and then used to simulate realistic data. We will assess performance using an example from five healthcare organizations that participate in the FDA Sentinel Initiative. The original evaluation compared the angiotensin-converting enzyme inhibitors (ACEI) to beta blockers (BB) with respect to the onset of angioedema, a potentially life-threatening allergic reaction.

B. METHODS

In **Section B.1** we introduce and detail the necessary notation. In **Section B.2** we describe several statistical techniques for using site-specific summary information to simulate realistic subject-level confounder, drug exposure, time-to-event and time-to-censoring data. **Section B.3** discusses the use of bootstrap sampling as a means of comparing data simulated using summary statistics to the underlying data source that was used to compute the summary statistics.

1. Notation

We assume that subject *i* (*i*=1,…,n), has a set of *J* binary covariates, $\boldsymbol{B}_i = (B_{i1}, \ldots, B_{iJ})$ and a set of K categorical covariates, $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{iK})$. Let B_{ij} represent the *j*th binary covariate (*j*=1,…,*J*) with levels 0 and 1 and C_{ik} represent the *k*th categorical covariate (*k*=1,…,*K*) with levels *l* (*l*=1,..,L_k). We denote the probability, or mean, of the *j*th binary covariate, P($B_j = 1$), as p_{B_j}, and the vector of probabilities (marginal proportions) for the L_k levels of the *k*th categorical variable as p_{C_k}. When referring to the combined vector of binary and categorical covariates for subject *i* we use the notation $Z_i = (B_i, C_i)$.

The binary indicator of exposure is denoted as X_i for the *i*th subject and equals 1 if they are exposed and 0 otherwise. Additionally, we assume that each subject has an exposure time T_i which is the minimum of their event time, T_i^E, and their censoring time, T_i^C. Y_i is the outcome and is coded 1 if an event occurred ($T_i^E \leq T_i^C$) and 0 otherwise. For the remainder of the manuscript we assume that the following observed data elements are available to inform simulation: *X*, ***Z***, *Y*, and *T*.

2. Simulating Data Using Summary Information

The general framework of this approach is to simulate subject level covariates, exposure variables, and time-to-event or censoring outcome variables from summary level data. This section will outline a set of summary estimates that can be obtained from collected data and used to simulate different types of subject-level data. In addition, this section will outline which summary information is needed to emulate the simulations that were conducted in this report, and how the summary information can be used to generate purely simulated subject-level data that resembles the observed data in predefined ways. The material presented here is by no means meant to be exhaustive. Indeed, there are a myriad of possibilities for modification and customization within the framework considered. The following list provides a brief summary of the methods considered and the statistics collected for each type of data that we want to simulate and is followed by sections which provide a brief description of each technique.

1. Covariates:
    a. Binned Multivariate Normal Distributions: Marginal empirical probabilities for each variable level and bivariate correlations between variables.
    b. Chain of regressions: Coefficients from a sequence of regression models predicting each covariate in-turn.
    c. Bootstrap: Samples with replacement from the set of observed subject covariate vectors.
2. Exposure given Covariates:
    a. Bernoulli distribution: Probability of exposure dependent on covariates predicted by model with coefficients from pre-specified, site-specific propensity score models estimated via logistic regression.
    b. Bootstrap: Samples with replacement from the set of observed subject covariate vectors including drug exposure.
3. Time-to-Outcome given exposure and covariates:
    a. Parametric survival regression: Time-to-Event simulated from Weibull distributions where parameters are conditional on covariates/exposure and are estimated with site-specific models.
    b. Bootstrap: Samples with replacement from subject-level covariate, exposure, exposure time and outcome vectors.
4. Time-to-censoring given exposure and covariates:
    a. Parametric survival regression: Time-to-Censoring simulated from Weibull distributions where parameters are conditional on covariates/exposure and are estimated with site-specific models, under three different scenarios:
        i. Censoring not dependent on covariates;
        ii. Censoring not dependent on covariates, but censoring times come from a mixture of a discrete distribution and a continuous distribution.
        iii. Censoring depends on covariates and drug exposure.
    b. Bootstrap: Samples with replacement from subject-level covariate, drug exposure, exposure time and censoring vectors.

a. Covariates

i. Multivariate Normal Thresholding

Techniques for simulating correlated binary and ordinal categorical variables using multivariate normal thresholding, referred to as the "mean mapping method" , have been described previously in working papers by Leisch, Weingessel and Hornik(23) and Kaiser, Träger and Leisch(35), respectively. The authors also developed two R packages, bindata and orddata, that include routines for generating correlated binary and categorical variables. These packages include multiple methods for simulating data, including multivariate normal thresholding as one option.

The examples presented in this report make use exclusively of bindata and orddata but there are also several other publications and R packages implementing simulation of correlated binary, ordinal, normal continuous, and non-normal continuous random variables, as well as combinations of the aforementioned. For SAS, multivariate normal thresholding for simulation of correlated ordinal random variables is discussed and implemented in SAS/IML by Wicklin (2003)(37).

The following sections provide a brief overview of the selection of methods that were used to generate the results presented in this report.

Binary and Categorical Covariates

Simulation of binary and categorical variables that reflect the correlation observed in the source data can be accomplished by taking random draws from the multivariate normal distribution with a specified correlation matrix and then thresholding (binning) each of the simulated normal variables at quantiles that correspond to steps in the estimated cumulative distribution function of each binary and/or categorical covariate. In this report, the R packages bindata and ordata, both of which implement multivariate normal thresholding, were used to generate binary covariates and/or combinations of binary and categorical covariates, respectively. Both packages use a similar methodology wherein observed data dependencies, or summary statistics, are connected to the multivariate normal distribution by creating equivalence between the pairwise correlations of the observed binary and/or categorical covariates and the correlation matrix of a bivariate normal distribution. Explicit detail can be found in the publications of Kaiser et al,(35) but in brief, the method equates the pairwise joint empirical cumulative distribution function estimated from the data, with a standard bivariate normal distribution. Kaiser, Träger and Leisch(35) show that

$$\sum_{\substack{1\le c_1\le L_1-1\\ 1\le c_2\le L_2-1}} \mathrm{F}_{C_1C_2}(c_1,c_2) = \rho_{C_1C_2}\sqrt{\sigma_{C_1}^2}\sqrt{\sigma_{C_2}^2} - \mu_{C_1}\mu_{C_2} - L_1L_2 + L_1\sum_{c_2=1}^{L_2-1}\mathrm{F}_{C_2}(c_2) + L_2\sum_{c_1=1}^{L_1-1}\mathrm{F}_{C_1}(c_1)$$

where $\mathrm{F}_{C_1C_2}(c_1,c_2) = P(C_1 \le c_1, C_2 \le c_2)$ and $\rho_{C_1C_2}$ is the Spearman correlation between C_1 and C_2.

The bivariate normal distribution with unknown correlation $\rho_{Z_1Z_2}$ is substituted on the left-hand side to yield

$$\sum_{\substack{1\le c_1\le L_1-1\\ 1\le c_2\le L_2-1}} \Phi_{Z_1Z_2}\left(q_{F_{C_1}(c_1)}, q_{F_{C_2}(c_2)}, \rho_{Z_1Z_2}\right)$$

$$= \rho_{C_1C_2}\sqrt{\sigma_{C_1}^2}\sqrt{\sigma_{C_2}^2} - \mu_{C_1}\mu_{C_2} - L_1L_2 + L_1\sum_{c_2=1}^{L_2-1}F_{C_2}(c_2) + L_2\sum_{c_1=1}^{L_1-1}F_{C_1}(c_1)$$

where $F_{C_i}(c)$ is the estimated marginal cumulative distribution function of an integer-valued discrete random variable C_i. The marginal cumulative distribution function is defined here as $F_{C_i}(c) = \sum_{k\le c} P(C_i = k)$. The values $q_{F_{C_i}(c)}$ for $c \in \{1,\dots,L_i\}$ are the quantiles of a standard normal distribution that correspond to the $F_{C_i}(c)^{\text{th}}$ percentiles of the observed variables. L_1 and L_2 are the number of levels of C_1 and C_2. (μ_{C_1}, μ_{C_2}) , $(\sigma_{C_1}^2, \sigma_{C_2}^2)$, and $\rho_{C_1C_2}$ are the means, variances, and Pearson correlation of C_1 and C_2, computed as if the variables were continuous.

To estimate the multivariate normal correlation parameter $\rho_{Z_1Z_2}$ in equation above, the mean and variance for each variable is computed as $\hat{\mu}_{C_i} = \sum_{k=1}^{L_i} k\,\hat{P}(C_i = k)$ and $\hat{\sigma}_{C_i}^2 = \sum_{k=1}^{L_i}(k-\mu_{C_i})^2\,\hat{P}(C_i = k)$. The correlation $\rho_{C_iC_j}$, where $i \ne j$ is estimated either by the correlation matrix computed in the observed data, or assembled from the observed marginal and pairwise common probabilities:

$$\hat{\rho}_{C_1C_2} = \left[\sum_{c_i}\sum_{c_j} c_i c_j \hat{P}(C_1 = c_i, C_2 = c_j) - \sum_{c_i} c_i \hat{P}(C_1 = c_i)\sum_{c_j} c_j \hat{P}(C_2 = c_j)\right]$$

where $c_i, c_j \in \{1, \ldots, L_1\} \times \{1, \ldots, L_2\}$. The estimated bivariate standard normal correlation $\rho_{Z_1 Z_2}$ can be obtained via a root finding algorithm, or, as is done in the R package orddata, the function can be evaluated on a grid of points, and the value of $\rho_{Z_1 Z_2}$ estimated via interpolation. As previously stated, once an estimate of $\rho_{Z_1 Z_2}$ is obtained, it is generally straight-forward to simulate data from the bivariate standard normal distribution with correlation $\rho_{Z_1 Z_2}$ and categorize the simulated variables at the corresponding set of quantiles $\{q_{F_{C_i}(c)} : c \in \{1, \ldots, L_i\}\}$.

To simulate multiple variables from multiple data partner sites, an analyst prepares a program to calculate the above univariate and bivariate summary statistics and deploys the program independently at each data partner. Once all of the individual and pairwise summary statistics have been computed, returned and reviewed, the pairwise correlation estimates are combined into a single correlation matrix, Σ^{sim}, and used to simulate *K* variables, z_i, where i=1,…,K, from the multivariate standard normal distribution. Each z_i is then binned in the following way to form the categorical variable c_i^{sim}.

$$c_i^{sim} = \begin{cases} 1 & if\ z_i \epsilon [0, q_{F_{C_i}(1)}) \\ 2 & if\ z_i \epsilon [q_{F_{C_i}(1)}, q_{F_{C_i}(2)}) \\ & \vdots \\ L_i & if\ z_i \epsilon [q_{F_{C_i}(L_i - 1)}, q_{F_{C_i}(L_i)}] \end{cases}$$

ii. Chains of Regressions

Chains of regressions can be used to capture information about both marginal and conditional distributions of binary and categorical covariates, and are simple to fit using most available statistical software packages. In the section below, we describe a simple algorithm we implemented for fitting a sequence of logistic regressions for predicting all binary covariates, and a sequence of multinomial logistic regressions for predicting categorical covariates. The sequence of regression models captures multivariate conditional relationships between binary and categorical covariates. The parameters summarizing models fit to the original subject level data can in turn be used to simulate subject level data from either the binomial or multinomial distributions.

Binary Covariates

The equations below show a simple chain of logistic regressions using a set of binary covariates. The regressions can be fit in any order and we suggest starting with the simpler models with fewer covariates then ordering the covariates, $\boldsymbol{B}$ in ascending order by their estimated marginal means (equivalent to the proportion). This ordering will yield better results in cases where some of the binary covariates have means that are close to zero. Letting α_{mn} represent each of the parameters in models with the nth variable as the outcome, where m=0,…,J-1 and n=1,…,J, the regression chain would go as follows:

$$logit(B_{i1}) = \alpha_{01}$$
$$logit(B_{i2}) = \alpha_{02} + \alpha_{12}B_{i1}$$
$$logit(B_{i3}) = \alpha_{03} + \alpha_{13}B_{i1} + \alpha_{23}B_{i2}$$

.
.
.

$$logit(B_{ij}) = \alpha_{0J} + \sum_{m=1}^{J-1} \alpha_{mJ}B_{im}$$

In a setting where multiple data partners are contributing data, estimates of all the $\boldsymbol{\alpha_{mn}}$ are saved and returned from each data partner. As described below, an analyst can then use these estimates to simulate (subject level) binary covariates with similar characteristics.

Categorical Covariates

For categorical covariates with more than two levels, we need information about the marginal distribution of each covariate and the pairwise associations between covariates. For a set of K categorical variables C_k $(k = 1, \dots, K)$ with L_k levels in indexed by l, where $l = 1, \dots, L_k$, the marginal distribution is determined by the marginal probabilities $P(C_k = l)$. A simple way to quantify the dependence of C_k on other covariates is to fit a multinomial logistic regression (note the multinom function in the nnet package in R was applied in the simulation) to C_k conditional on all other covariates. To operationalize this, we fit multinomial logistic regressions to each categorical variable in a chained fashion as we cycle through the K categorical covariates. If present, binary covariates are included as predictors in each model, and categorical covariates used as outcomes are successively added as predictors in each subsequent model. When fitting multinomial logistic regressions we obtain an estimated linear predictor for each level of the categorical outcome variable. As an example, for two categorical variables C_1 and C_2 with levels $L_1 = 4$ and $L_2 = 5$, respectively, and allowing the first level to be the reference category, we would fit a chain of two multinomial logistic regressions as follows. Letting C_{ik} represent the ith subject's value for the kth covariate when have

$$mlogit(C_{i1}|B_{i1}, \dots, B_{iJ}) = \begin{cases} level\ 2 & \gamma_{01}^2 + \sum_{j=1}^{J} \beta_{j1}^2 B_{ij} \\ level\ 3 & \gamma_{01}^3 + \sum_{j=1}^{J} \beta_{j1}^3 B_{ij} \\ level\ 4 & \gamma_{01}^4 + \sum_{j=1}^{J_B} \beta_{j1}^4 B_{ij} \end{cases}$$

and

$$mlogit\left(C_{i2}|C_{i1},B_{i1},\dots,B_{iJ}\right)=\begin{cases} level\ 2 & \gamma_{02}^{2}+\boldsymbol{\gamma}_{12}^{2}C_{i1}^{l}+\sum_{j=1}^{J}\beta_{j2}^{2}B_{ij} \\ level\ 3 & \gamma_{02}^{3}+\boldsymbol{\gamma}_{12}^{3}\boldsymbol{C}_{i1}^{l}+\sum_{j=1}^{J}\beta_{j2}^{3}B_{ij} \\ level\ 4 & \gamma_{02}^{4}+\boldsymbol{\gamma}_{12}^{4}\boldsymbol{C}_{i1}^{l}+\sum_{j=1}^{J}\beta_{j2}^{4}B_{ij} \\ level\ 5 & \gamma_{02}^{5}+\boldsymbol{\gamma}_{12}^{5}\boldsymbol{C}_{i1}^{l}+\sum_{j=1}^{J}\beta_{j2}^{5}B_{ij} \end{cases}$$

where $\boldsymbol{C}_{i1}^{l}$ is (L_1-1)x1 vector of indicator variables representing C_{i1} equal to level l, where $l=1,\dots,L_k$. γ_{0k}^{l} is the intercept and β_{jk}^{l} is the coefficient for binary confounder j for the l^{th} level of the multinomial model for the k^{th} categorical confounder. For the second model we further specify that $\boldsymbol{\gamma}_{12}^{l}$ is a 1x(L_1-1) vector of coefficients relating the first categorical variable in the chain to C_{i2}. The model coefficients are returned to be used in data simulation.

Note that a similar approach could be used to generate summary information for continuous covariates. However, since in the current application we have no need to include continuous covariates we will not cover the topic here.

After estimates for all α and β coefficients have been estimated and returned, an analyst can simulate a set of correlated binary and categorical covariates of any sample size by looping through all of the variables in the same order that the parameter estimates were generated and simulating subject-level values as follows:

$$B_{i1}^{sim}\sim Bernoulli(P(B_{i1}))$$
$$B_{i2}^{sim}\sim Bernoulli(P(B_{i2}|B_{i1}))$$

.
.
.

$$B_{iJ}^{sim}\sim Bernoulli(P(B_{ij}|B_{i1},\dots,B_{iJ-1}))$$
$$C_{i1}^{sim}\sim Multinomial(1,P(C_{i1}|B_{i1},\dots,B_{iJ-1}))$$
$$C_{i2}^{sim}\sim Multinomial(1,\boldsymbol{P}(C_{i2}|C_{i1},B_{i1},\dots,B_{iJ-1}))$$

.
.
.

$$C_{iK}^{sim}\sim Multinomial(1,\boldsymbol{P}(C_{iK}|C_{i1},\dots,C_{iK-1},B_{i1},\dots,B_{iJ-1}))$$

b. Exposure and Propensity Score Model

To compute summary statistics which will allow simulation of a dichotomous exposure that depends on measured variables, we fit a propensity score model, i.e., the probability of exposure given confounders, using logistic regression with the exposure variable as the outcome regressed on a function of the binary and categorical covariates. This model can be as flexible as required – perhaps including higher order terms for single covariates, interactions between covariates, splines, etc. – and would typically include only terms that are thought to be related either to both the outcome and the exposure, though variables related only to the outcome can also be useful. The propensity score model takes the form

$$logit\big(P(X_i|\boldsymbol{Z}_i)\big) = \boldsymbol{\theta}'\boldsymbol{f}(\boldsymbol{Z}_i),$$

where $\boldsymbol{Z}_i$ is a vector of covariates for subject *i*, $\boldsymbol{\theta}$ is a vector of coefficients and $\boldsymbol{f}(\boldsymbol{Z}_i)$ is a vector-valued function of the observed covariates which may include interactions, polynomial functions and/or regressions splines. The coefficients are retained and can be used by an analyst to compute the probability of receiving treatment based on covariates. Once estimates of $\boldsymbol{\theta}$ are returned from all data partners subject-level exposure values, X_i^{sim}, can be generated as

$$X_i^{sim} \sim Bernoulli(P(X_i|Z_i))$$

c. Time-To-Event

A basic summary of event and censoring times observed in the source data can be obtained by fitting flexible parametric survival regressions for time-to-event and time-to-censoring. Here we briefly provide the basic details of time-to-event data and connect them to the Weibull form of the parametric survival model. The presentation is intended to give insight into the techniques that were used in this report and to serve as a brief reference for an analyst who intends to conduct similar simulations.

The Weibull distribution for survival times, t, is defined by scale parameter λ and shape parameter γ and has probability density function $f(t) = \lambda^{\gamma}\gamma t^{\gamma-1}e^{-\lambda t^{\gamma}}$. The survival function can be defined in terms of the density function as

$$S(t) = 1 - \int_0^t f(u)du = e^{-\lambda^{\gamma}t^{\gamma}}.$$

The hazard function is related to the survival function in the following way

$$\begin{aligned} h(t) &= \frac{d}{dt}\log(S(t)) \\ &= \frac{f(t)}{S(t)} \\ &= \lambda^{\gamma}\gamma t^{\gamma-1}. \end{aligned}$$

The cumulative hazard function can be written in terms of the hazard function or the survival function

$$H(t) = \int_0^t h(u)du, \text{ or}$$

$$H(t) = -\log\big(S(t)\big) = \lambda^{\gamma}t^{\gamma}.$$

To fit a model to observed survival data assuming a Weibull distribution with scale parameter λ_i^E and shape parameter γ^E, where the superscript *E* indicates that the model is for the outcome event as

opposed to a censoring event (denoted with superscript *C* below), i.e., the event times, T_i^E, are distributed as $Weibull(\lambda_i^E, \gamma^E)$, and we typically fit the following model of the form

$$\log(T_i^E) = \eta_0 + \eta_x X_i + \boldsymbol{\eta}' \boldsymbol{Z}_i + \sigma W_i,$$

where W_i follows the extreme value distribution. Parameter estimates for this model can be obtained via maximum likelihood using pre-existing functions in most modern statistical packages, including R, SAS and Stata. If R or SAS is used for estimation, to estimate the hazard, survival and cumulative hazard functions detailed above, we transform the coefficients and scale parameter from the Weibull regression model to yield the Weibull scale and shape parameters $\lambda_i^E = e^{-(\eta_0 + \eta_x X_i + \boldsymbol{\eta}' \boldsymbol{Z}_i)}$ and $\gamma^E = \frac{1}{\sigma}$. In this case we have chosen to estimate a constant shape parameter for all subjects, which can be thought of as assuming that the distribution of event times across drug exposures and covariate groups has the same shape, but that the conditional hazard, λ_i^E, varies yielding longer or shorter times depending on the observed covariate values. We could also allow σ to vary by covariate strata, yielding stratum-specific shape parameters $\gamma_i^E = \frac{1}{\sigma_i}$.

These parameter estimates can be used to simulate event times either by directly simulating using a pre-existing function for generating random Weibull variables in existing statistical software, or by simulating random values from a uniform distribution on the interval [0,1] and using the probability integral transform as outlined in Bender et al (2005)(28). For example, using the methodology discuss in Bender et al, we can generate Weibull random variables T_i^{Esim} as

$$T_i^{Esim} = \left(-\log(U_i) e^{(\eta_0 + \eta_x X_i + \boldsymbol{\eta}' \boldsymbol{Z}_i)/\sigma}\right)^{\sigma},$$

where U_i is a random draw from a uniform distribution on the interval [0,1]. The next section describes similar models for estimating parameters to define several potential distributions for time-to-censoring, T_i^{Csim}. Regardless of the censoring distribution used, when the parameters estimates from each data partner are returned, an analyst can simulate T_i^{Esim} and T_i^{Csim} and define the "observed" follow-up time, T_i^{sim}, and event indicator, Y_i^{sim}, for each simulation as

$$T_i^{sim} = \min(T_i^{Esim}, T_i^{Csim})$$

and

$$Y_i^{sim} = \begin{cases} 0 & if\ T_i^{Esim} \leq T_i^{Csim}, \\ 1 & otherwise. \end{cases}$$

d. Time-to-Censoring

We followed a similar methodology for modeling censoring times as was described in the section above. We considered three different models for censoring times: i) simple, covariate independent, ii) simple, covariate independent allowing for common discrete prescription lengths and iii) covariate adjusted.

i. Simple Independent Censoring

For Simple Independent Censoring, we assumed that the censoring times followed a Weibull distribution with shape parameter λ^C and scale parameter γ^C, both independent of drug exposure and covariates. Parameter estimates were obtained via maximum likelihood by fitting the following model for time-to-censoring, T_i^C,

$$\log(T_i^C) = \nu_0 + \sigma^C W_i,$$

yielding $\lambda^C = e^{-\nu_0/\sigma^C}$ and $\gamma^C = \frac{1}{\sigma^C}$.

ii. Simple and Discrete Independent Censoring

In some cases, there may be additional features of the observed censoring distribution that we wish to mimic in our simulations. For example, there may be a high frequency of particular censoring times like 30, 60 or 90 days reflecting common prescription lengths for certain medications. These additional features can be modeled as discrete times and combined with the continuous censoring distribution based on frequency of occurrence. In this case we can add an additional step to the modeling process for censoring times. Instead of fitting the Weibull model to all of the censoring times, we first estimate the probability of being censored at particular times, e.g. 30, 60 or 90 days, and then fit the Weibull time-to-censoring model to the data excluding these times. When simulating data, prior to generating a censoring time, we first take a draw from a multinomial distribution with a bin for each of the unique times that were removed and one additional bin that indicates that the time should come from the continuous distribution. Given the example mentioned above where 30, 60 and 90-day prescriptions are notably more frequent in the context of the overall distribution of censoring times, we would draw a random variable M_i from the multinomial distribution with probability vector υ given by

$$\upsilon = \left(P\left(T_i^C = 30\right), P\left(T_i^C = 60\right), P\left(T_i^C = 0\right), 1 - \sum_{t \in \{30,60,90\}} P\left(T_i^C = t\right) \right)'$$

If M_i takes on realizations $m_i \in \{1, 2, 3, 4\}$ we adhere to the following rule for assigning censoring times:

$$T_i^{Csim} = \begin{cases} 30, & if\ m_i = 1 \\ 60, & if\ m_i = 2 \\ 90, & if\ m_i = 3 \\ Weibull(\lambda^*, \gamma^*), & if\ m_i = 4, \end{cases}$$

where λ^* and γ^* are the shape and scale parameters from a time-to-censoring model fit to the data with censoring times of 30, 60 and 90 days excluded.

iii. Covariate and Exposure Dependent Censoring

Another approach to flexibly model the censoring time is to allow it to depend on drug exposure and covariates. Common survival model analysis techniques, such as Cox's proportional hazards model, assume that the censoring distribution is independent of the event distribution given covariates. In practice censoring distributions are often highly related to exposures and covariates. For example, older adults may be more likely to stop taking medication due to other comorbidities. Another important example is when comparing newly marketed drugs to drugs that have been on the market for some time (exposure of interest compared to a comparator). The new drugs are often more expensive and may initially be prescribed in 30-day intervals while a comparator drug that may be on the market longer may initially be prescribed in 90-day intervals. Therefore, censoring is more likely to occur earlier for older adults and those on newer drugs. To allow for this complexity we assume a similar Weibull time-to-censoring model as was detailed above for events:

$$\log\left(T_i^C\right) = \nu_0 + \nu_x X_i + \boldsymbol{\nu}' \boldsymbol{Z}_i + \sigma^C W_i,$$

yielding $\lambda_i^C = e^{-(\nu_0 + \nu_x X_i + \boldsymbol{\nu}' \boldsymbol{Z}_i)/\sigma^C}$ and $\gamma^C = \frac{1}{\sigma^C}$. Some statistical approaches may be able to account for this type of covariate and exposure dependent censoring (e.g. Cox PH regression approaches which

adjust for the covariates and exposure in the model directly), however other approaches may or may not be able to handle this assumption as well (e.g. approaches which do not directly condition on both covariates and exposure in the model as is the case with some propensity score based approaches). Therefore, allowing for data to be simulated with this flexibility may be important. In our simulation study in **Section IV**, we will compare these three censoring approaches (simple independent censoring, discrete and simple independent censoring, and covariate and exposure dependent censoring) to assess performance. Note that other censoring mechanisms such as those that combine discrete censoring and covariate and/or exposure dependent censoring are also easily implemented in our current simulation framework. For simplicity, we only present the three general approaches since they cover the most common scenarios typically observed in our data setting. Ultimately, we did not observe large differences between censoring approaches adding more nuanced methods.

e. Simulate Site Data Given Summary Statistics

Given the summary information from each site, the process of data simulation for a given site follows a similar sequence to that used to collect summary statistics. Specifically if you simulate covariates using the multivariate normal thresholding approach you first begin with the matrix of common probabilities or correlation matrix and means as outlined in **Section II.B.2.a.i** from your dataset. Given this data you calculate the normal distribution $\rho_{Z_1 Z_2}$ and quantile cut-offs $\{q_{F_C(c)}: c \in \{1, \dots, L\}\}$. Then simulate continuous covariates from a multivariate normal distributed with correlation $\rho_{Z_1 Z_2}$ and use the quantile cut-offs to derive the simulated binary and categorical covariates.

After simulating binary and categorical covariates, the site-specific propensity score model is used to simulate exposure, X_i, which is generated from a Bernoulli distribution with probability $P(X_i) = (1 + e^{-\theta' Z_i})^{-1}$. After simulating the binary, categorical and exposure variables, we can then simulate corresponding event and censoring times using the parameters from the parametric survival models that were fit at each site. For each subject record, we simulate both an event and a censoring time. Given these two times we take the minimum, i.e., $T_i^{sim} = \min(T_i^{Esim}, T_i^{Csim})$. If $T_i^{Esim} \leq T_i^{Csim}$ then the event indicator, Y_i, equals 1, and otherwise $Y_i = 0$. In **Section II.C.1** we will detail the summary information and simulation process for a specific data example.

When actually conducting a simulation evaluation one would like not only to mimic actual data, but also change certain parameters of interest such as the strength of the relationship between the exposure of interest and outcome. This would typically be done by maintaining observed associations between covariates and outcome in the dataset and simply changing η_x to the desired log(Hazard Ratio) comparing exposed to unexposed. Since data is simulated within site, if the interest is in assessing performance of methods when site heterogeneity exists, then η_x must be different for each site, i.e., an interaction between treatment and site must be created.

Given the summary information the researcher has the ability to change any parameter of interest to explore a large range of questions easily (e.g. 1: methods performance for varying confounding relationships (change the propensity model coefficients or covariate outcome coefficients), 2: methods performance when creating site heterogeneity (differentially changing the relationship between exposure and confounder across sites), or 3: methods performance when missing confounders in the model dependent on confounder prevalence (miss-specify the model dropping different confounders from the method to see if it makes a difference)). Therefore, this new approach allows one to mimic realistic data situations, but does not constrain the types of questions to be asked.

3. Bootstrap as Gold Standard

Since the purpose of this exercise is to be able to simulate data that mirrors real data, one way to evaluate the effectiveness of the proposed simulation methods is to compare them to the empirical distribution of the data that we are attempting to mimic. We can accomplish this by using bootstrap sampling from the subject level data as a kind of gold standard. Comparisons can take place on a number of dimensions but will focus on marginal and pairwise correlations.

C. ASSESSMENT OF DIFFERENT DATA SIMULATION APPROACHES

In this simulation assessment, we compare data generated using bootstrap sampling of subject-level covariates, drug exposures, follow-up times and indications of angioedema to our proposed approach of using summary information only. In what follows we detail the test data used and the summary statistics related to each type of data and compare the results of analyses performed on simulated data with those obtained from analyzing parallel bootstrap samples.

1. Example Data

Data for this example was taken from a cohort study conducted within Sentinel to evaluate the relative risk of angioedema between users of ACE-inhibitors (ACEI) and a select group of beta-blocker (BB) users.

Table 1. Sample characteristics by site (n=150,000)

Variable	Site 1	Site 2	Site 3	Site 4	Site 5
Sample Size	48127	19275	33399	45012	4187
Age					
30-44 (Ref)	28.1	22.3	9.5	22	25.4
45-54	28.5	24.3	11.8	24.3	30
55-64	26.7	24.9	15.5	25.8	31.8
≥65	16.7	28.4	63.2	27.9	12.8
Sex					
Male (Ref)	50.8	48.8	48.6	51.1	51
Female	49.2	51.2	51.4	48.9	49
Comorbidity Score					
≤ 0 (Ref)	78.9	77.5	63.9	75.7	77.2
1	21.1	22.5	36.1	24.3	22.8
Emergency Visits					
0 (Ref)	80.7	80.9	84.4	87.1	78.8
≥ 1	19.3	19.1	15.6	12.9	21.2
Hospital Stays					
0 (Ref)	89.9	90.3	84	85.6	89.3
≥ 1	10.1	9.7	16	14.4	10.7
Year					
2008	13.2	26.5	21.5	23.5	21.5
2009	27	22.7	21.2	21.8	22.2
2010	23.1	19.3	19.4	19.5	19.2
2011	19	16.3	18.8	18.4	19.2
2012	17.8	15.2	19.2	16.8	17.8

Approximately 2.5 million records were available from the original inquiry, representing five data partners – referred to in what follows as sites 1-5. In all, the data included information on 1.4 million ACEI users (56%) and 1.1 million BB users. The proportional data contributions of each of the five sites was 32%, 13%, 23%, 30%, and 2.5% of the total sample, respectively. To ensure that meaningful results could be generated within a reasonable time frame, we reduced the computational burden by taking a simple random sample of 6% of the 2.5 million records, or 150,000 subjects, as a test cohort. This randomly sampled cohort included subject-level covariate values, drug exposures, follow up times and indications of angioedema or censoring. The sample comprises 84,351 ACE-inhibitor users and 65,649 beta-blocker users and includes three binary covariates (sex; number of emergency visits (EVs) in the last 180 days: 0, 1+; and number of inpatient hospital stays (HSs) in the last 180 days: 0, 1+) and three categorical covariates (Age: 18-44, 45-54, 55-64, 65-99; Comorbidity Index: ≤ 1, >1; and Year: 2008, 2009, 2010, 2011, 2012). Summaries of covariate distributions by site are presented in **Table 1.**

In addition to site, **Table 2** further stratifies the sample by drug exposure and contains sample sizes, counts of angioedema events, rates of angioedema per thousand person-years and average follow-up days.

Table 2. Sample size, average* person-days of follow-up, number of events and event rates per 1000 person-years by site and drug exposure (n=150,000)

Site	Drug	N	Average Person Time	Events	Rates
SITE 1	BB	21080	106.9	11	1.78
	ACEI	27047	128.1	51	5.38
SITE 2	BB	8435	151.7	4	1.14
	ACEI	10840	180.5	22	4.11
SITE 3	BB	13881	130.7	6	1.21
	ACEI	19518	154.2	35	4.25
SITE 4	BB	20358	118.6	11	1.66
	ACEI	24654	144.5	40	4.10
SITE 5	BB	1895	116	0	0.00
	ACEI	2292	147.5	6	6.48

*Average is sum of person-time across all subjects divided by number of subjects N

Overall, there were 186 angioedema events in the test sample with unadjusted rates of 4.6 per 1,000 and 1.5 per 1,000 person-years among ACEI users and BB users, respectively. Variation in the rate of angioedema across data partner sites was small, with the exception of site 5 where only six events in total were sampled, all of which were in the ACE-inhibitor group.

As described in **Section II.B** above, we cover two different ways of simulating correlated binary and categorical covariates: multivariate normal thresholding and regression chains. For multivariate normal thresholding, we need estimates of the common probabilities. For chain regressions, we need coefficients from sequential chains of logistic and multinomial regressions. The site-specific marginal probability estimates from **Table 1** can be used to estimate the means and the correlation can either be computed directly by treating the variables as continuous data, or calculated using common probabilities. As an example, **Table 3** shows the site-averaged matrix of common probabilities (reference level excluded for space). The pairwise Pearson correlation matrices of the covariates at each site are

shown in Appendix **Table B 1**. The computed model coefficients for chain of regressions are shown in **Table B 2** and **Table B 3**.

To generate drug exposure data, we use propensity score models fit separately at each data partner site to predict the probability of exposure given simulated covariates, generating an exposure $X_i \sim Bernoulli\left(\left(1+e^{-Z_i\theta_s}\right)^{-1}\right)$ for each subject. In this example, we estimated a propensity score model for exposure to ACEIs versus BBs with the same functional form at each of the five sites. There are a variety of ways to increase the complexity of the propensity score models if needed, including fitting different models at each data partner, specifying higher order terms or interactions, or using data driven model-selection algorithms. Odds ratios, standard errors, and c-statistics from the site-specific propensity score models are displayed in **Table 4**.

Table 3. Observed probabilities* over data partner sites (n=150,000; reference levels excluded)

	1+ HS	1+ EV	1+ CS	SexF	45-54	55-64	65+	2009	2010	2011	2012
1+ HS	12.7	5.4	8.6	6.5	2.0	2.5	6.1	2.9	2.5	2.3	2.3
1+ EV	5.4	16.6	7.6	8.7	3.6	3.5	5.4	3.8	3.4	3.1	3.1
1+ CS	8.6	7.6	25.6	13.3	4.0	5.2	12.6	5.8	5.2	4.8	4.8
SexF	6.5	8.7	13.3	49.9	11.0	11.5	16.7	11.6	10.3	9.3	8.6
45-54	2.0	3.6	4.0	11.0	23.0	0.0	0.0	5.7	4.9	4.1	3.7
55-64	2.5	3.5	5.2	11.5	0.0	23.9	0.0	5.8	5.0	4.3	4.1
65+	6.1	5.4	12.6	16.7	0.0	0.0	31.8	6.8	6.3	6.1	6.0
2009	2.9	3.8	5.8	11.6	5.7	5.8	6.8	23.4	0.0	0.0	0.0
2010	2.5	3.4	5.2	10.3	4.9	5.0	6.3	0.0	20.6	0.0	0.0
2011	2.3	3.1	4.8	9.3	4.1	4.3	6.1	0.0	0.0	18.4	0.0
2012	2.3	3.1	4.8	8.6	3.7	4.1	6.0	0.0	0.0	0.0	17.5

*Probabilities in diagonal cells are frequency of variables (%). Probability in each off diagonal cell is the frequency (%) of co-occurrence of the corresponding row and column binary variables.

Table 4. Observed odds ratios from site-specific propensity score models

	Site 1		Site 2		Site 3		Site 4		Site 5	
	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE
Intercept	1.39	0.044	1.46	0.063	1.46	0.064	1.39	0.041	1.24	0.12
Age										
45-54	1.64	0.041	1.88	0.083	1.6	0.079	1.58	0.045	1.95	0.171
55-64	1.67	0.043	1.68	0.073	1.56	0.073	1.53	0.043	1.94	0.168
65+	1.28	0.038	1.31	0.056	1.29	0.051	1.26	0.036	1.76	0.198
1+ CS	0.53	0.013	0.56	0.022	0.65	0.017	0.54	0.014	0.54	0.045
1+ EV	0.8	0.021	0.74	0.032	0.8	0.025	0.84	0.025	0.67	0.059
1+ HS	0.54	0.02	0.52	0.031	0.5	0.016	0.51	0.016	0.5	0.064
SexF	0.61	0.012	0.64	0.019	0.83	0.019	0.64	0.013	0.56	0.037
Year										
2009	1.16	0.037	1.04	0.045	1.14	0.04	1.08	0.031	1.11	0.109
2010	1.14	0.037	1.03	0.047	1.12	0.04	1.13	0.034	1.14	0.116
2011	1.09	0.037	0.96	0.045	1.06	0.038	1.12	0.034	1.13	0.116
2012	1.02	0.035	0.93	0.045	1.11	0.04	1.03	0.032	1.05	0.109
C-stat	0.64		0.64		0.61		0.64		0.66	

For time-to-event we utilize parametric survival models assuming an underlying Weibull distribution (discussed in **Section II.B**). The Weibull distribution is extremely flexible and can accommodate a wide variety of data. As with propensity score estimation there are a variety of choices that can be made when specifying models for the time-to-event outcomes. For simplicity, we have chosen to again fit models with the same functional form separately at each data partner. Hazard ratios, standard errors, and 95% confidence intervals from these models are tabulated in **Table 5**. For comparison with the more commonly used semi-parametric framework for analyzing survival data, **Table 6** displays results from fitting the standard Cox proportional hazards model to the time-to-angioedema data at each data partner.

Table 5. Observed Weibull hazard ratios, standard errors and 95% confidence interval estimates for time-to-angioedema by data partner (n=150,000)

	Site 1			Site 2			Site 3			Site 4			Site 5		
	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI
ACEI	3.3	1.1	(1.7, 6.6)	3.8	2.2	(1.2, 11.5)	3.8	1.7	(1.6, 9.3)	3.5	1.3	(1.7, 7.1)	Inf	Inf	Inf Inf
45-54	1.7	0.7	(0.8, 3.6)	0.4	0.2	(0.1, 1.2)	0.6	0.5	(0.1, 3.2)	0.7	0.3	(0.3, 1.6)	0.7	1.0	(0.0, 11.6)
55-64	1.6	0.6	(0.7, 3.4)	0.8	0.4	(0.3, 2.1)	0.8	0.6	(0.2, 3.3)	0.6	0.3	(0.3, 1.4)	1.6	1.9	(0.2, 16.0)
65+	1.1	0.5	(0.4, 2.7)	0.4	0.3	(0.1, 1.4)	1.1	0.7	(0.3, 3.8)	0.6	0.3	(0.3, 1.4)	0.9	1.4	(0.1, 16.0)
1+ CS	1.9	0.6	(1.0, 3.4)	1.3	0.7	(0.5, 3.7)	1.5	0.5	(0.8, 3.0)	1.4	0.5	(0.7, 2.7)	10.2	9.5	(1.7, 62.5)
1+ EV	1.2	0.4	(0.6, 2.5)	1.6	0.9	(0.6, 4.5)	1.4	0.6	(0.6, 3.1)	1.4	0.5	(0.7, 2.9)	2.3	2.4	(0.3, 18.8)
1+ HS	0.4	0.3	(0.1, 1.6)	0.4	0.5	(0.0, 3.6)	0.6	0.3	(0.2, 1.8)	2.2	0.8	(1.0, 4.5)	0.7	1.0	(0.0, 11.2)
SexF	1.5	0.4	(0.9, 2.5)	0.7	0.3	(0.3, 1.6)	0.7	0.2	(0.4, 1.3)	2.0	0.6	(1.1, 3.5)	0.7	0.6	(0.1, 3.9)
2009	2.3	1.2	(0.8, 6.7)	0.5	0.3	(0.2, 1.7)	1.2	0.6	(0.4, 3.2)	1.6	0.8	(0.6, 4.2)	Inf	Inf	Inf Inf
2010	1.6	0.9	(0.5, 4.9)	1.1	0.5	(0.4, 2.8)	0.7	0.4	(0.2, 2.3)	2.1	1.0	(0.8, 5.3)	Inf	Inf	Inf Inf
2011	1.9	1.1	(0.6, 5.9)	0.4	0.3	(0.1, 1.7)	1.7	0.8	(0.7, 4.4)	1.7	0.9	(0.7, 4.6)	0.0	0.0	(0.0, 0.0)
2012	3.0	1.7	(1.0, 9.0)	0.8	0.5	(0.2, 2.6)	1.1	0.6	(0.4, 3.1)	2.0	1.0	(0.7, 5.2)	88.3	Inf	Inf Inf

Table 6. Observed Cox PH hazard ratios, standard errors and 95% confidence interval estimates for time-to-angioedema by data partner

	Site 1			Site 2			Site 3			Site 4			Site 5		
	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI
ACEI	3.4	1.2	(1.7, 6.6)	3.8	2.1	(1.3, 11.4)	3.8	1.7	(1.6, 9.2)	3.5	1.2	(1.8, 7.0)	Inf	Inf	(0.0, Inf
45-54	1.7	0.7	(0.8, 3.6)	0.4	0.2	(0.1, 1.2)	0.6	0.5	(0.1, 3.2)	0.8	0.3	(0.3, 1.7)	0.8	1.1	(0.0, 12.6)
55-64	1.6	0.6	(0.8, 3.5)	0.8	0.4	(0.3, 2.2)	0.8	0.6	(0.2, 3.3)	0.7	0.3	(0.3, 1.5)	1.7	2.0	(0.2, 17.3)
65+	1.1	0.5	(0.4, 2.7)	0.5	0.3	(0.1, 1.4)	1.1	0.7	(0.3, 3.8)	0.7	0.3	(0.3, 1.4)	1.1	1.7	(0.1, 19.6)
1+ CS	1.9	0.6	(1.0, 3.4)	1.3	0.7	(0.5, 3.7)	1.5	0.5	(0.8, 3.0)	1.4	0.5	(0.7, 2.7)	9.0	8.2	(1.5, 53.3)
1+ EV	1.2	0.4	(0.6, 2.5)	1.6	0.8	(0.6, 4.5)	1.4	0.6	(0.6, 3.1)	1.4	0.5	(0.7, 2.9)	2.3	2.4	(0.3, 18.3)
1+ HS	0.4	0.3	(0.1, 1.5)	0.4	0.5	(0.0, 3.6)	0.6	0.3	(0.2, 1.8)	2.1	0.8	(1.0, 4.4)	0.8	1.1	(0.1, 11.8)
SexF	1.5	0.4	(0.9, 2.5)	0.7	0.3	(0.3, 1.6)	0.7	0.2	(0.4, 1.3)	2.0	0.6	(1.1, 3.5)	0.7	0.6	(0.1, 4.0)
2009	2.3	1.2	(0.8, 6.7)	0.5	0.3	(0.2, 1.7)	1.2	0.6	(0.4, 3.2)	1.6	0.8	(0.6, 4.2)	Inf	Inf	(0.0, Inf
2010	1.6	0.9	(0.5, 4.9)	1.1	0.5	(0.4, 2.8)	0.7	0.4	(0.2, 2.3)	2.1	1.0	(0.8, 5.3)	Inf	Inf	(0.0, Inf
2011	1.9	1.1	(0.6, 5.9)	0.4	0.3	(0.1, 1.7)	1.7	0.8	(0.7, 4.4)	1.7	0.9	(0.7, 4.6)	0.9	Inf	(0.0, Inf
2012	3.0	1.7	(1.0, 9.0)	0.8	0.5	(0.2, 2.6)	1.1	0.6	(0.4, 3.1)	2.0	1.0	(0.7, 5.1)	1.0	Inf	(0.0, Inf

With respect to the censoring distribution in the sample data, we explore three different formulations. In the first, and most basic scenario we assume that each subject's censoring time follows the exact same distribution regardless of their covariate values or exposure status (Simple Independent Censoring). Our summary information for this case comes from site-specific Weibull survival models with censoring as the outcome and only an intercept and a scale parameter specified. The results from fitting these models at each site are shown in **Table 7**. The second censoring model also assumes simple censoring but allows simulated data to reflect a small number of extremely common prescription times, e.g. 30 or 90 days (Discrete and Simple Independent Censoring). In this case, we specified that the three most common exposure times, or modes of the treatment period distribution, were 44, 104 and 365 days. These times correspond to prescriptions of 30 and 90 days with a 14-day post-exposure allowance and the administrative censoring time of 365 days (**Table 8**). As described in the methods section we then implemented a two-step sampling scheme, where for each subject a draw is first taken from a multinomial distribution with the result indicating that the censoring time should either be one of the three most common times or should be sampled from the simple independent censoring model. Parameters from fitting the discrete and simple independent censoring model to data where the three most common episode lengths are removed is shown in **Table 9**. The third censoring model allows censoring to depend on the same set of covariates and exposures (Covariate and Exposure Dependent Censoring). Results from the site-specific, covariate and exposure adjusted Weibull time-to-censoring models are shown in **Table 10**.

Table 7. Observed intercept and scale term from site-specific simple independent censoring models (site-specific Weilbull time-to-censoring model)

	Site 1	Site 2	Site 3	Site 4	Site 5
Intercept	4.83	5.22	5.03	4.94	4.95
Scale	0.88	0.68	0.85	0.87	0.87

Table 8. Three most common observed follow up times in days and corresponding proportion of all censoring times by data partner

Site 1		Site 2		Site 3		Site 4		Site 5	
Days	%	Days	%	Days	%	Days	%	Days	%
44	33.6	44	5.7	44	25.1	44	30.4	44	29.7
104	5.8	114	37.5	104	9.5	104	5.9	104	4.8
365	10.7	365	18.6	365	16.6	365	14.3	365	14.3

Table 9. Observed intercept and scale term from site-specific simple and discrete censoring model* with common times removed

	Site 1	Site 2	Site 3	Site 4	Site 5
Intercept	4.85	5.04	4.94	4.90	4.90
Scale	0.78	0.78	0.77	0.76	0.75

*Note: A two-step censoring model: (1) estimate the probability of being censored at 44, 104, and 365 (values shown in Table 8), and (2) fit the site-specific Weibull time-to-censoring model to the data excluding these times.

Table 10. Observed hazard ratios, standard errors and 95% confidence intervals from site-specific Weibull time-to censoring model conditional on covariates

	Site 1			Site 2			Site 3			Site 4			Site 5		
	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI	HR	SE	95% CI
ACEI	0.83	0.01	(0.82, 0.85)	0.82	0.01	(0.80, 0.85)	0.85	0.01	(0.83, 0.86)	0.81	0.01	(0.80, 0.83)	0.81	0.03	(0.76, 0.86)
1+ HS	1.00	0.02	(0.97, 1.04)	1.04	0.03	(0.98, 1.10)	1.08	0.02	(1.05, 1.12)	1.05	0.02	(1.02, 1.08)	1.16	0.07	(1.03, 1.30)
1+ EV	1.07	0.01	(1.04, 1.10)	1.06	0.02	(1.01, 1.10)	1.09	0.02	(1.06, 1.13)	1.09	0.02	(1.06, 1.12)	0.97	0.04	(0.89, 1.06)
1+ CS	1.03	0.01	(1.01, 1.06)	1.05	0.02	(1.02, 1.09)	1.04	0.01	(1.01, 1.06)	1.01	0.01	(0.99, 1.04)	1.04	0.04	(0.96, 1.13)
SexF	1.03	0.01	(1.01, 1.05)	1.01	0.01	(0.99, 1.04)	1.00	0.01	(0.98, 1.03)	1.02	0.01	(1.00, 1.04)	1.03	0.03	(0.97, 1.10)
Age															
45-54	0.81	0.01	(0.79, 0.83)	0.83	0.02	(0.80, 0.87)	0.83	0.02	(0.79, 0.87)	0.82	0.01	(0.80, 0.84)	0.82	0.03	(0.76, 0.90)
55-64	0.74	0.01	(0.72, 0.76)	0.75	0.02	(0.72, 0.79)	0.74	0.02	(0.71, 0.78)	0.76	0.01	(0.74, 0.78)	0.72	0.03	(0.66, 0.78)
65+	0.66	0.01	(0.64, 0.68)	0.72	0.02	(0.69, 0.75)	0.67	0.01	(0.65, 0.70)	0.69	0.01	(0.67, 0.71)	0.76	0.04	(0.68, 0.84)
Year															
2009	1.02	0.02	(0.99, 1.05)	0.99	0.02	(0.95, 1.03)	0.95	0.02	(0.92, 0.98)	0.99	0.01	(0.96, 1.01)	0.94	0.04	(0.86, 1.03)
2010	1.00	0.02	(0.97, 1.03)	0.99	0.02	(0.95, 1.04)	0.93	0.02	(0.90, 0.96)	0.96	0.01	(0.93, 0.99)	0.94	0.05	(0.86, 1.04)
2011	0.99	0.02	(0.96, 1.02)	0.96	0.02	(0.92, 1.01)	0.89	0.02	(0.86, 0.92)	0.96	0.01	(0.94, 0.99)	0.94	0.05	(0.85, 1.03)
2012	0.97	0.02	(0.94, 1.00)	1.04	0.02	(1.00, 1.09)	0.89	0.02	(0.86, 0.92)	0.97	0.01	(0.94, 1.00)	1.43	0.07	(1.30, 1.58)

2. Performance of Covariate Generation Procedures

As an informal way of comparing covariates generated using multivariate normal thresholding or chain regression with those obtained from subject-level bootstrap samples, we summarize the means and standard errors of the simulation distributions of the pooled-data common probabilities in **Table 11, Table 12,** and

Table 13 (observed probability matrix shown in **Table 3**). Inspecting these tables, we can see that the marginal probabilities and the common probabilities are in very good agreement across data generation methods.

Table 11. Simulation probabilities* using multivariate normal thresholding (n=150,000, 5,000 simulations)

	1+ HS		1+ EV		1+ CS		SexF		45-54		55-64		65+		2009		2010		2011		2012	
	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE
1+ HS	12.70	(0.09)	5.40	(0.06)	8.60	(0.07)	6.50	(0.06)	2.40	(0.04)	3.00	(0.04)	5.60	(0.06)	2.90	(0.04)	2.60	(0.04)	2.30	(0.04)	2.20	(0.04)
1+ EV	5.40	(0.06)	16.60	(0.10)	7.60	(0.07)	8.80	(0.07)	3.90	(0.05)	4.00	(0.05)	4.90	(0.06)	3.90	(0.05)	3.50	(0.05)	3.10	(0.04)	3.10	(0.04)
1+ CS	8.60	(0.07)	7.60	(0.07)	25.60	(0.11)	13.30	(0.09)	4.70	(0.05)	6.00	(0.06)	11.80	(0.08)	5.80	(0.06)	5.20	(0.06)	4.80	(0.06)	4.80	(0.06)
SexF	6.50	(0.06)	8.80	(0.07)	13.30	(0.09)	49.90	(0.13)	11.30	(0.08)	11.90	(0.08)	16.40	(0.09)	11.70	(0.08)	10.30	(0.08)	9.20	(0.07)	8.70	(0.07)
45-54	2.40	(0.04)	3.90	(0.05)	4.70	(0.05)	11.30	(0.08)	23.00	(0.11)	0.00	(0.00)	0.00	(0.00)	5.50	(0.06)	4.80	(0.06)	4.20	(0.05)	3.90	(0.05)
55-64	3.00	(0.04)	4.00	(0.05)	6.00	(0.06)	11.90	(0.08)	0.00	(0.00)	23.90	(0.11)	0.00	(0.00)	5.70	(0.06)	5.00	(0.06)	4.40	(0.05)	4.10	(0.05)
65+	5.60	(0.06)	4.90	(0.06)	11.80	(0.08)	16.40	(0.09)	0.00	(0.00)	0.00	(0.00)	31.80	(0.11)	7.10	(0.06)	6.40	(0.06)	6.00	(0.06)	6.00	(0.06)
2009	2.90	(0.04)	3.90	(0.05)	5.80	(0.06)	11.70	(0.08)	5.50	(0.06)	5.70	(0.06)	7.10	(0.06)	23.40	(0.11)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
2010	2.60	(0.04)	3.50	(0.05)	5.20	(0.06)	10.30	(0.08)	4.80	(0.06)	5.00	(0.06)	6.40	(0.06)	0.00	(0.00)	20.60	(0.10)	0.00	(0.00)	0.00	(0.00)
2011	2.30	(0.04)	3.10	(0.04)	4.80	(0.06)	9.20	(0.07)	4.20	(0.05)	4.40	(0.05)	6.00	(0.06)	0.00	(0.00)	0.00	(0.00)	18.40	(0.10)	0.00	(0.00)
2012	2.20	(0.04)	3.10	(0.04)	4.80	(0.06)	8.70	(0.07)	3.90	(0.05)	4.10	(0.05)	6.00	(0.06)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	17.50	(0.10)

*In diagonal cells, % is the frequency of variables averaged over all simulations, SE is the standard errors of these frequencies across all simulations. In off-diagonal cells, % is P the frequency (%) of co-occurrence of the corresponding row and column binary variables averaged over all simulations, SE is the standard error of these frequencies across all simulations.

Table 12. Simulation probabilities* using regression chains (n=150,000, 5,000 simulations)

	1+ HS		1+ EV		1+ CS		SexF		45-54		55-64		65+		2009		2010		2011		2012	
	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE
1+ HS	12.70	(0.09)	5.40	(0.06)	8.60	(0.07)	6.50	(0.06)	2.00	(0.04)	2.50	(0.04)	6.10	(0.06)	2.90	(0.04)	2.50	(0.04)	2.30	(0.04)	2.30	(0.04)
1+ EV	5.40	(0.06)	16.60	(0.10)	7.60	(0.07)	8.80	(0.07)	3.60	(0.05)	3.50	(0.05)	5.40	(0.06)	3.80	(0.05)	3.40	(0.05)	3.10	(0.04)	3.10	(0.04)
1+ CS	8.60	(0.07)	7.60	(0.07)	25.60	(0.11)	13.30	(0.09)	4.00	(0.05)	5.20	(0.06)	12.60	(0.08)	5.80	(0.06)	5.20	(0.06)	4.80	(0.06)	4.80	(0.05)
SexF	6.50	(0.06)	8.80	(0.07)	13.30	(0.09)	49.90	(0.13)	11.00	(0.08)	11.50	(0.08)	16.70	(0.09)	11.60	(0.08)	10.30	(0.08)	9.30	(0.07)	8.60	(0.07)
45-54	2.00	(0.04)	3.60	(0.05)	4.00	(0.05)	11.00	(0.08)	23.00	(0.11)	0.00	(0.00)	0.00	(0.00)	5.70	(0.06)	4.90	(0.06)	4.10	(0.05)	3.70	(0.05)
55-64	2.50	(0.04)	3.50	(0.05)	5.20	(0.06)	11.50	(0.08)	0.00	(0.00)	23.90	(0.11)	0.00	(0.00)	5.80	(0.06)	5.00	(0.06)	4.30	(0.05)	4.10	(0.05)
65+	6.10	(0.06)	5.40	(0.06)	12.60	(0.08)	16.70	(0.09)	0.00	(0.00)	0.00	(0.00)	31.80	(0.11)	6.80	(0.07)	6.30	(0.06)	6.10	(0.06)	6.00	(0.06)
2009	2.90	(0.04)	3.80	(0.05)	5.80	(0.06)	11.60	(0.08)	5.70	(0.06)	5.80	(0.06)	6.80	(0.07)	23.40	(0.11)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
2010	2.50	(0.04)	3.40	(0.05)	5.20	(0.06)	10.30	(0.08)	4.90	(0.06)	5.00	(0.06)	6.30	(0.06)	0.00	(0.00)	20.60	(0.10)	0.00	(0.00)	0.00	(0.00)
2011	2.30	(0.04)	3.10	(0.04)	4.80	(0.06)	9.30	(0.07)	4.10	(0.05)	4.30	(0.05)	6.10	(0.06)	0.00	(0.00)	0.00	(0.00)	18.40	(0.10)	0.00	(0.00)
2012	2.30	(0.04)	3.10	(0.04)	4.80	(0.05)	8.60	(0.07)	3.70	(0.05)	4.10	(0.05)	6.00	(0.06)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	17.50	(0.10)

*In diagonal cells, % is the frequency of variables averaged over all simulations, SE is the standard errors of these frequencies across all simulations. In off-diagonal cells, % is P the frequency (%) of co-occurrence of the corresponding row and column binary variables averaged over all simulations, SE is the standard error of these frequencies across all simulations.

Table 13. Probabilities in datasets simulated using bootstrap sampling (n=150,000, 5,000 simulations)

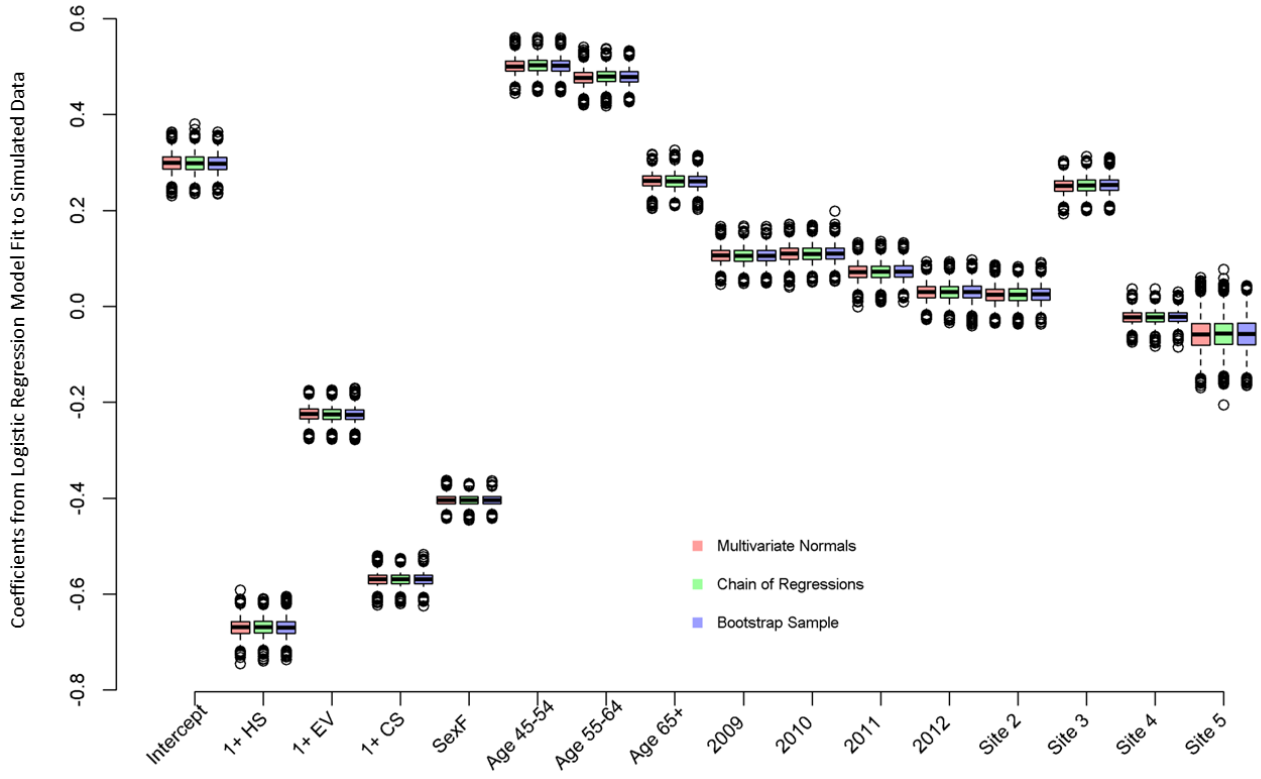
	1+ HS		1+ EV		1+ CS		SexF		45-54		55-64		65+		2009		2010		2011		2012	
	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE
1+ HS	12.70	(0.08)	5.40	(0.06)	8.60	(0.07)	6.50	(0.06)	2.00	(0.04)	2.50	(0.04)	6.10	(0.06)	2.90	(0.04)	2.50	(0.04)	2.30	(0.04)	2.30	(0.04)
1+ EV	5.40	(0.06)	16.60	(0.10)	7.60	(0.07)	8.80	(0.07)	3.60	(0.05)	3.50	(0.05)	5.40	(0.06)	3.80	(0.05)	3.40	(0.05)	3.10	(0.04)	3.10	(0.05)
1+ CS	8.60	(0.07)	7.60	(0.07)	25.60	(0.11)	13.30	(0.09)	4.10	(0.05)	5.20	(0.06)	12.60	(0.08)	5.80	(0.06)	5.20	(0.06)	4.80	(0.06)	4.80	(0.05)
SexF	6.50	(0.06)	8.80	(0.07)	13.30	(0.09)	49.90	(0.13)	11.00	(0.08)	11.50	(0.08)	16.70	(0.09)	11.60	(0.08)	10.30	(0.08)	9.30	(0.08)	8.60	(0.07)
45-54	2.00	(0.04)	3.60	(0.05)	4.10	(0.05)	11.00	(0.08)	23.00	(0.11)	0.00	(0.00)	0.00	(0.00)	5.70	(0.06)	4.90	(0.06)	4.10	(0.05)	3.70	(0.05)
55-64	2.50	(0.04)	3.50	(0.05)	5.20	(0.06)	11.50	(0.08)	0.00	(0.00)	23.90	(0.11)	0.00	(0.00)	5.80	(0.06)	5.00	(0.06)	4.30	(0.05)	4.10	(0.05)
65+	6.10	(0.06)	5.40	(0.06)	12.60	(0.08)	16.70	(0.09)	0.00	(0.00)	0.00	(0.00)	31.80	(0.11)	6.80	(0.06)	6.30	(0.06)	6.10	(0.06)	6.00	(0.06)
2009	2.90	(0.04)	3.80	(0.05)	5.80	(0.06)	11.60	(0.08)	5.70	(0.06)	5.80	(0.06)	6.80	(0.06)	23.40	(0.11)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
2010	2.50	(0.04)	3.40	(0.05)	5.20	(0.06)	10.30	(0.08)	4.90	(0.06)	5.00	(0.06)	6.30	(0.06)	0.00	(0.00)	20.60	(0.10)	0.00	(0.00)	0.00	(0.00)
2011	2.30	(0.04)	3.10	(0.04)	4.80	(0.06)	9.30	(0.08)	4.10	(0.05)	4.30	(0.05)	6.10	(0.06)	0.00	(0.00)	0.00	(0.00)	18.40	(0.10)	0.00	(0.00)
2012	2.30	(0.04)	3.10	(0.05)	4.80	(0.05)	8.60	(0.07)	3.70	(0.05)	4.10	(0.05)	6.00	(0.06)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	17.50	(0.10)

*In diagonal cells, % is the frequency of variables averaged over all simulations, SE is the standard errors of these frequencies across all simulations. In off-diagonal cells, % is P the frequency (%) of co-occurrence of the corresponding row and column binary variables averaged over all simulations, SE is the standard error of these frequencies across all simulations.

3. Performance of Exposure given Covariate Procedures

Figure 2 provides boxplot summaries of the simulation distributions of the coefficients from the pooled-data propensity score models. The figure shows excellent agreement across covariate generation methods. Site-specific propensity score coefficients are shown in Appendix Figures B 1 a-e.

Figure 2. Simulation distributions of coefficients from pooled data propensity score model (5,000 simulations)



4. Performance of Outcome Given Exposure and Covariates Generation Procedures

Figure 3, Figure 4 and **Figure 5** show boxplot summaries of the simulation distributions of the coefficients from a Cox PH model fit to each simulated dataset pooled over site. Specifically, the Cox PH model fit was a model adjusting for indicator of ACEI exposure, all covariates, and site indicator variables in a single model. Across data generation methods the coefficient distributions show very good agreement, with the exception of the bootstrap distribution for the dummy indicator for site 5. Thus, to make the figures viewable very extreme outliers from that site were not plotted. The difference in the coefficient distribution for site 5 for the bootstrap appears to be due to the parametric model potentially not fitting the data very well, or from an alternate perspective, to the data not being particularly suitable for the estimation of hazard ratios. This can be seen more clearly in the site-specific figures included as Appendix **Figure B 3 a-e** and **Figure B 4 a-e** where the distributions for some of the coefficients under the parametric models at site 5 do not generate nearly as many extreme values. In practice, we may choose to exclude a data partner like site 5 because there were no events in the BB group in the source data which led to instability in the coefficient estimates from that site.

The choice of censoring model appears to have made little difference in the results. **Figure 6** and **Figure 7** show boxplot summaries of model coefficients compared across censoring methods holding the covariate data generation method constant.

Figure 3. Distribution of fitted coefficients from Cox PH outcome models to simulations with simple censoring (5,000 simulations)

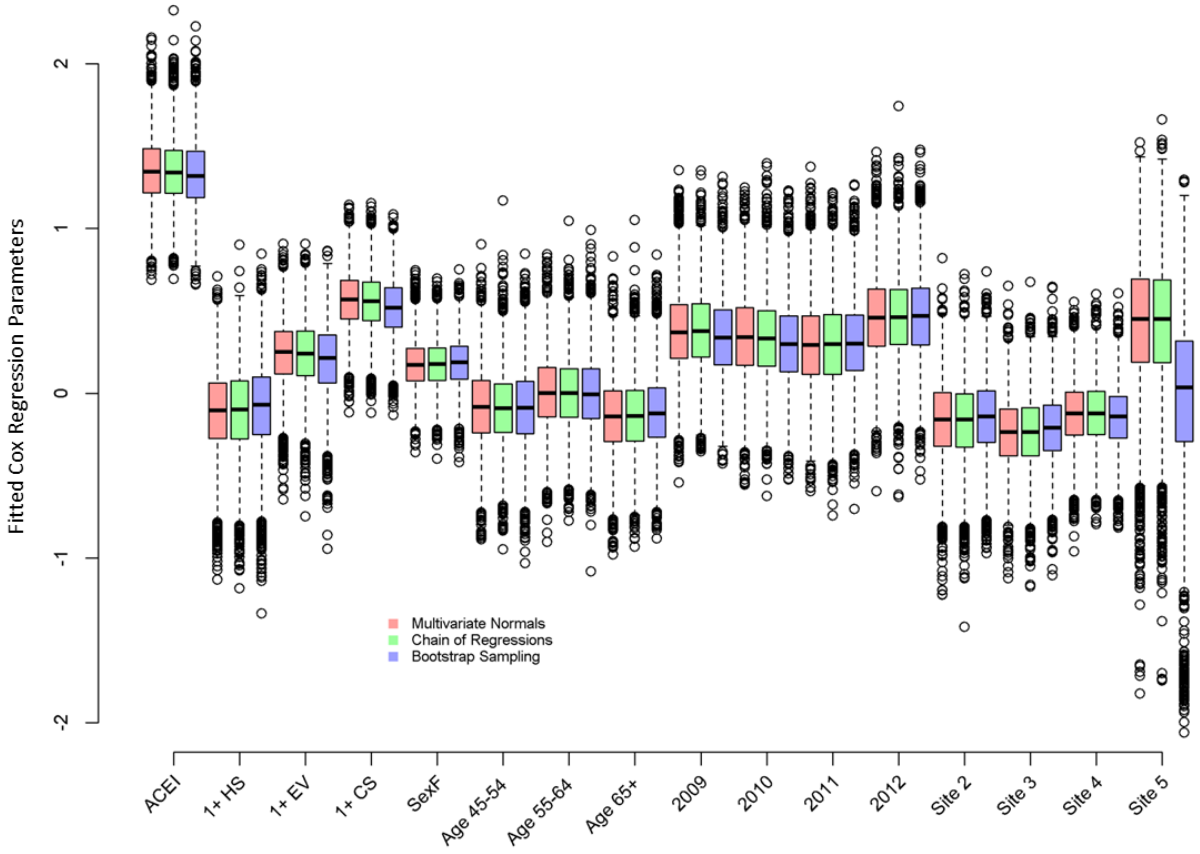


Figure 4. Distributions of fitted coefficients from Cox PH outcome model to simulations with simple censoring (5,000 simulations)

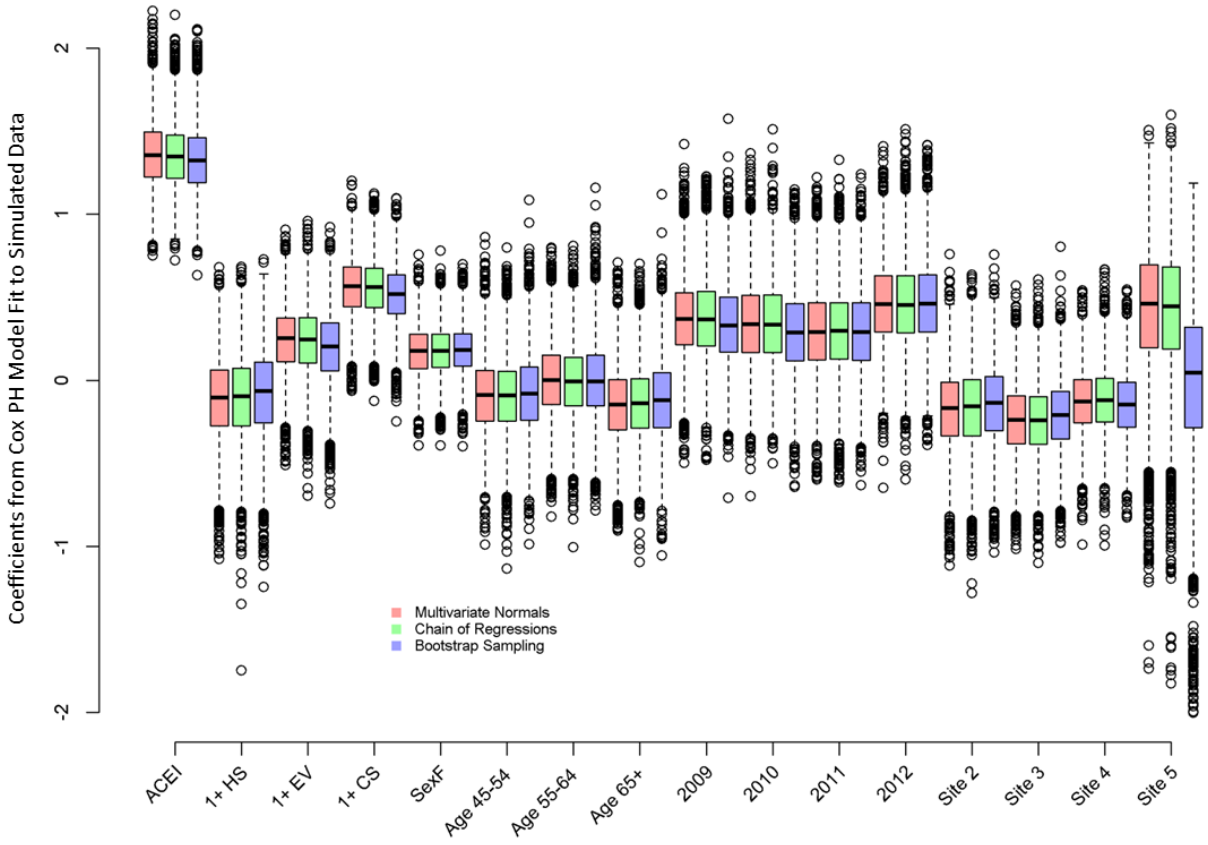


Figure 5. Distributions of fitted coefficients from Cox PH outcome model to simulations with covariate adjusted censoring (5,000 simulations)

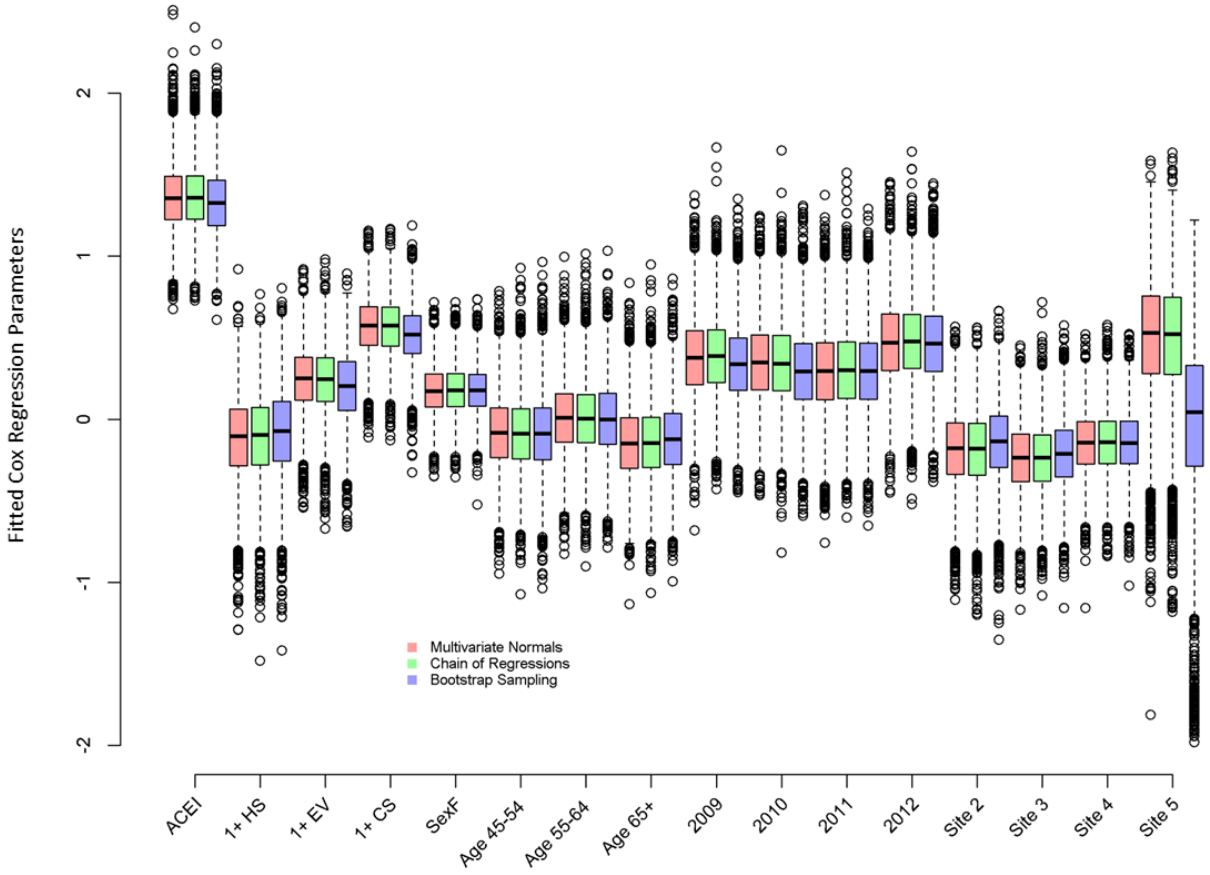


Figure 6. Distribution of coefficients from Cox PH outcome model fitted to simulations with different censoring models and multivariate normal thresholding (5,000 simulations)

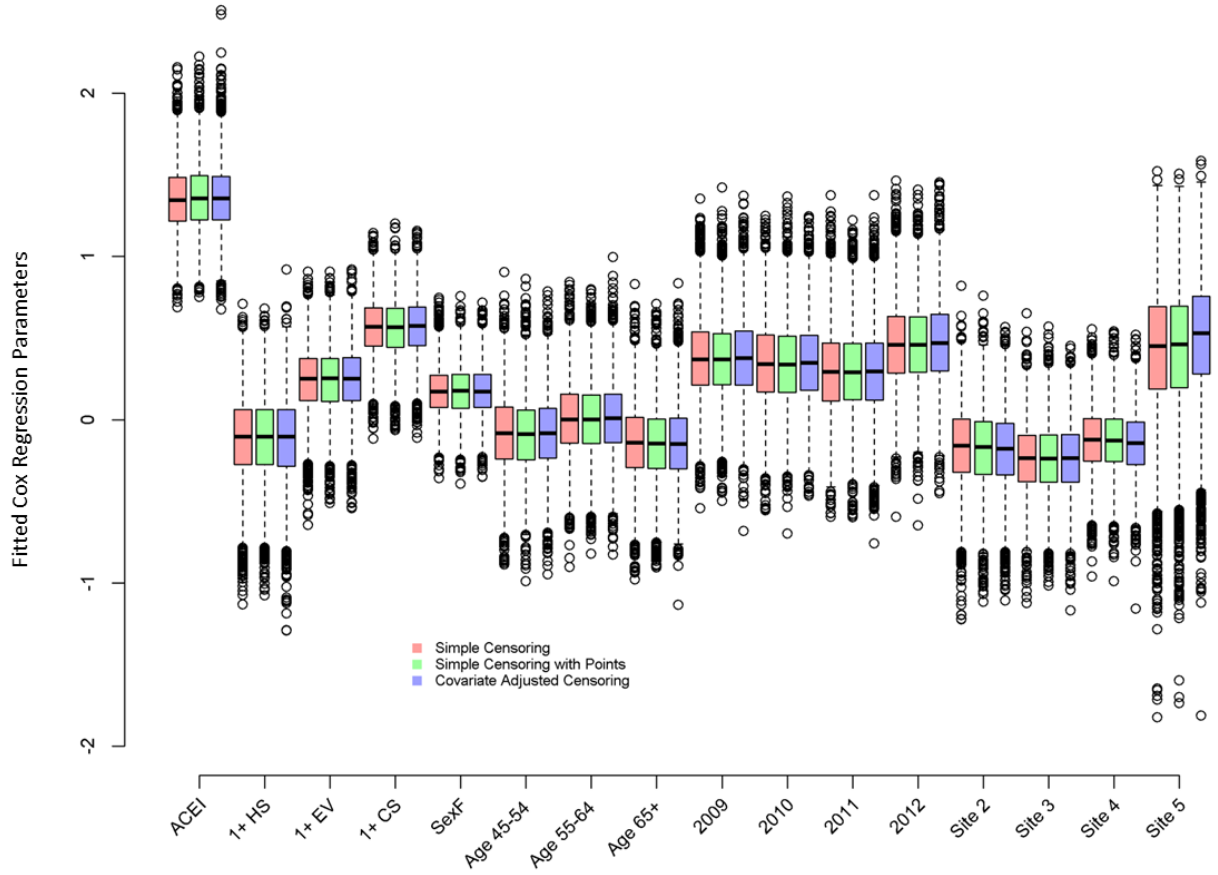
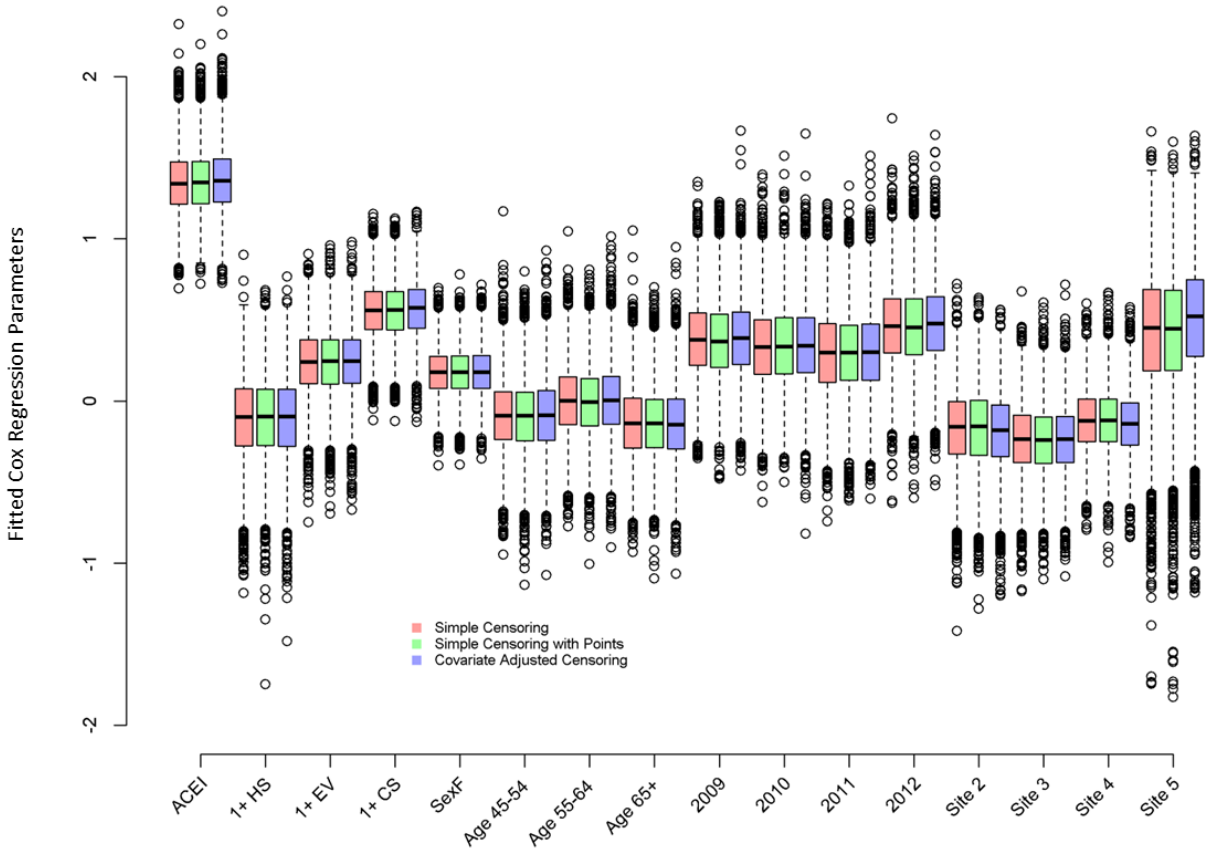


Figure 7. Distribution of coefficients from Cox PH outcome model fitted to simulations with different censoring models and using regressions chains (5,000 simulations)



D. DISCUSSION

Section II has presented a simple simulation approach for generating subject level survival outcome data from summary level information. This is an advantage in situations where sharing subject level data is not possible. Because the summary level information is from real data, our simulation approach mimics realistic real-world data examples that can be used for future method evaluation studies. It keeps the main features of complex data intact, in this case marginal distributions and pairwise correlations, and requires minimal summary information to conduct the simulation study. All proposed simulation methods in this approach share a key feature of maintaining complex confounder relationships including correlation between confounders. Certain methods may break down when there are strong interdependent confounder relationships and this would be something one would want to evaluate when assessing method performance. Another key feature of this approach is that it allows for complex relationships between the censoring mechanism and covariates. Often in simulation studies a very simple censoring assumption is made in which censoring is not reliant on other information such as covariates. In practice, censoring is strongly related to covariate information including the exposure of interest. For example, people who are older may be more likely to stop taking medications or switch to a new medication due to drug interactions given that they tend to take more medications overall, yielding age-dependent censoring. Further, like the example presented in this paper, censoring can also depend

on the exposure of interest. Those that received beta blockers were more likely to be given a 30-day prescription versus ACEI users who were more likely to receive a 90 day. When comparing the performance of statistical through simulation studies, retaining these key features allows an analyst to, say, narrow the focus of the simulation evaluation to datasets that resemble the data that could be sampled directly at a particular data network, or to mimic data coming from the different data partners in a data network to evaluate statistical methods meant to be deployed in a distributed fashion.

We use this simulation approach in our simulation study presented in **Section IV** comparing methods for estimating HR using the ACEI and Angioedema example.

III. STATISTICAL METHODS FOR THE NON-DISTRIBUTED DATA SETTING

We will present several conditional survival regression methods that are typically applied to datasets in which subject level data could be shared across sites without concerns about patient privacy or concerns that the data are proprietary. Conditional methods can be defined as methods that condition directly on confounders (adjusted confounder methods) or condition on strata.

A. COX PH REGRESSION ADJUSTING FOR CONFOUNDERS

Assume at site *s* (*s* = 1, … , *S*), we observe data from participant *i* (*i* = 1,… , n_s) that has either received the exposure of interest, $X_{si} = 1$, or the comparator, $X_{si} = 0$. Furthermore, each participant has a set of baseline confounders, $\boldsymbol{Z}_{si}$, δ_{si} indicating whether they have experienced the outcome before the end of the study follow-up period and 0 otherwise and T_{si} for time to event or censoring. Further, define a set of site indicator variables $\boldsymbol{S}_{si} = (S_{si}^2, \ldots S_{si}^S)$ where S_{si}^j is 1 if s=j and 0 otherwise.

Consider an adjusted Cox's PH regression model adjusting for confounders and site indicator variables (Adj Confounders+Site),

$$\lambda(T_{si}, \delta_{si} | X_{si}, \boldsymbol{Z}_{si}, \boldsymbol{S}_{si}) = \lambda_0(T_{si}) exp\left[\beta_X^{Adj} X_{si} + \boldsymbol{\beta}_z \, \boldsymbol{Z}_{si} + \boldsymbol{\beta}_S \, \boldsymbol{S}_{si}\right]. \qquad (1)$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, β_X^{Adj} is the log(HR) comparing the exposure of interest to the comparator, $\boldsymbol{\beta}_z$ is a $1 \times p$ vector of unknown confounder regression parameters, and $\boldsymbol{\beta}_S$ is a $1 \times (S-1)$ vector of unknown site regression parameters. We would estimate the regression model using standard partial maximum likelihood estimation to derive the fitted estimates $\hat{\beta}_x$, $\widehat{\boldsymbol{\beta}}_z$, and $\widehat{\boldsymbol{\beta}}_S$.

For a given analysis time we would be interested in assessing the following hypothesis: $H_O: \beta_X = 0$ versus $H_A: \beta_X > 0$. To assess this hypothesis, we would derive a test statistic. One standard test statistic is the Wald test statistic, $\hat{\beta}_X / \sqrt{\hat{V}(\hat{\beta}_X)}$. However, it is more common to form a score test statistic (a.k.a. Log Rank Statistic) since it is relatively more powerful, while still being straightforward to calculate. The corresponding Log Rank test statistic is:

$$LR = \frac{\sum_{\{s,i:\delta_{si}=1\}}\left[X_{si} - \frac{\sum_{\{k,l:T_{kl}\geq T_{si}\}} X_{kl}\exp\left(\widehat{\boldsymbol{\beta}}_z^{(0)}\boldsymbol{Z}_{kl} + \widehat{\boldsymbol{\beta}}_S^{(0)}\boldsymbol{S}_{kl}\right)}{\sum_{\{k,l:T_{kl}\geq T_{si}\}}\exp\left(\widehat{\boldsymbol{\beta}}_z^{(0)}\boldsymbol{Z}_{kl} + \widehat{\boldsymbol{\beta}}_S^{(0)}\boldsymbol{S}_{kl}\right)}\right]}{\sqrt{\sum_{\{s,i:\delta_{si}=1\}}\left[\frac{\sum_{\{k,l:T_{kl}\geq T_{si}\}} X_{kl}\exp\left(\widehat{\boldsymbol{\beta}}_z^{(0)}\boldsymbol{Z}_{kl} + \widehat{\boldsymbol{\beta}}_S^{(0)}\boldsymbol{S}_{kl}\right)}{\sum_{\{k,l:T_{kl}\geq T_{si}\}}\exp\left(\widehat{\boldsymbol{\beta}}_z^{(0)}\boldsymbol{Z}_{kl} + \widehat{\boldsymbol{\beta}}_S^{(0)}\boldsymbol{S}_{kl}\right)} - \left(\frac{\sum_{\{k,l:T_{kl}\geq T_{si}\}} X_{kl}\exp\left(\widehat{\boldsymbol{\beta}}_z^{(0)}\boldsymbol{Z}_{kl} + \widehat{\boldsymbol{\beta}}_S^{(0)}\boldsymbol{S}_{kl}\right)}{\sum_{\{k,l:T_{kl}\geq T_{si}\}}\exp\left(\widehat{\boldsymbol{\beta}}_z^{(0)}\boldsymbol{Z}_{kl} + \widehat{\boldsymbol{\beta}}_S^{(0)}\boldsymbol{S}_{kl}\right)}\right)^2\right]}}$$

where $\widehat{\boldsymbol{\beta}}_z^{(0)}$ and $\widehat{\boldsymbol{\beta}}_S^{(0)}$ are the fitted parameter estimates of model (1) under H_O that $\beta_X = 0$. Large positive values of *LR* signify that the exposure of interest has a higher hazard ratio compared to a comparator.

Direct adjustment for categorical confounders performs well if the number of confounders is low relative to the number of outcomes observed. For the FDA Sentinel setting our datasets tend to have a large sample size, but we are often in the rare event setting when the number of outcomes is still relatively small. Therefore, as the number of confounders increases we may not be able to directly adjust for all of the confounders. To address this issue propensity score methods have been proposed to account for confounding instead and we will describe several different approaches in the following sections.

B. COX PH REGRESSION ADJUSTING FOR PROPENSITY SCORES (LINEARLY, INDICATORS, OR B-SPLINES)

Propensity score methods are used to reduce the confounder information into a summary score to address large number of confounders in a more parsimonious model framework. We will outline three different propensity score approaches using Cox PH regression through adjustment.

The propensity score is defined as the probability of being exposed given a set of confounders. Specifically, for our setting we can define it dependent on baseline confounders and site assuming the following logistic model,

$$e_{si} = P(X_{si}|\boldsymbol{Z}_{si},\boldsymbol{S}_{si}) = exp\left[\boldsymbol{\gamma}_z\,\boldsymbol{Z}_{si} + \boldsymbol{\gamma}_S\,\boldsymbol{S}_{si}\right]/(1 + exp\left[\boldsymbol{\gamma}_z\,\boldsymbol{Z}_{si} + \boldsymbol{\gamma}_S\,\boldsymbol{S}_{si}\right]).$$

The estimated propensity score, $\hat{e}_{si}$, is derived fitting the logistic regression model using standard MLE theory to obtain regression parameter estimates of $\boldsymbol{\gamma}_z$ and $\boldsymbol{\gamma}_S$.

The most common approach to adjust for propensity score has been through a linear adjustment using the following Cox PH model,

$$\lambda(T_{si},\delta_{si}|X_{si},\boldsymbol{e}_{si}) = \lambda_0(T_{si})exp\left[\beta_X^{PSAdj}X_{si} + \beta_p\;e_{si}\right]. \quad (2)$$

However, this approach has shown to be biased for survival outcomes as well as for continuous and binary outcomes likely due to residual confounding from model misspecification.(30) A more flexible approach to model the relationship between the propensity score and outcome is to use a set of propensity score indicator variables based on percentiles. Specifically, define a set of *K*-1 indicator variables $e_{si}^k = I\left[\left(e_{si} > e_{si}^{(100(k-1)/K)\%tile}\right) \cap \left(e_{si} \leq e_{si}^{(100k/K)\%tile}\right)\right]$ for *k*=2,…,*K*. Then include the indicator variables in the Cox PH regression model as follows,

$$\lambda\left(T_{si},\delta_{si}|X_{si},\boldsymbol{S}_{si},\boldsymbol{e}_{si}^{\cdot}\right) = \lambda_0(T_{si})exp\left[\beta_X^{PSIAdj}X_{si} + \boldsymbol{\beta}_p^I\boldsymbol{e}_{is}^I\right], \quad (3)$$

where $\boldsymbol{e}_{si}^{I} = (e_{si}^{2}, \ldots, e_{si}^{K})^{T}$ is a (*K*-1)x1 vector of propensity score indicator variables for site *s*. We call this method Cox PH "Adj PS Indicators". An open research question is how to choose *K* the number of propensity score strata which will depend on number of outcomes, strength of confounding, and the underlying distribution of the observed propensity scores. We will vary *K* in the simulation study including 5, 10, 15, and 20 strata.

The last approach that we will evaluate is to adjust for propensity scores using splines and in particular cubic b-splines with two internal knots at 33.3% and 66.6% quantiles yielding 5 parameters in the model. We use b-splines because they are computationally easy to fit. We will call this method Cox PH "Adj PS B-Splines" and it is fit using the following Cox PH model,

$$\lambda(T_{si}, \delta_{si} | X_{si}, \boldsymbol{S}_{si,} \boldsymbol{e}_{si}) = \lambda_0(T_{si}) exp\left[\beta_X^{PSBSAdj} X_{si} + \boldsymbol{\beta}_p^{bs} f(\boldsymbol{e}_{is})\right], \quad (4)$$

where $f(\boldsymbol{e}_{is})$ are 5 cubic b-spline basis functions.

C. COX PH REGRESSION ADJUSTING FOR SITE-SPECIFIC PROPENSITY SCORES (INDICATORS OR B-SPLINES)

Often the relationship between receiving the exposure of interest and confounders may be different at each site. Sites are healthcare plans across different regions throughout the US. An example of reasons for differences may be due to different formulary plans for dispensing of medications or the availability of a new vaccine. Therefore, you may not expect the uptake of a new medical product to be similar across healthcare plans and/or the relationship between confounders to be similar. If the uptake is different, but the relationship between uptake and confounders across sites is the same, then the previous propensity score model adjusting for site as a main effect is correctly specified. However, if the relationship between confounders and exposure is different across sites you may want to model site-specific propensity score models. An example may be that at certain sites the new vaccine was primarily given to children at 3 months, but other sites the vaccine was given across all infant ages. Further, when moving to the distributed data setting site-specific propensity score information may only be available since data cannot be combined across sites into a single dataset. Therefore, there are multiple reasons site-specific propensity scores could be used and there are different approaches to adjust for them.

The first approach, Cox PH "Adj Site-PS Indicators", we propose and will evaluate assumes the following Cox PH model with adjustment for site and site-specific propensity score indicator variables interacted with site,

$$\lambda(T_{si}, \delta_{si} | X_{si}, \boldsymbol{S}_{si,} \boldsymbol{e}_{si}^{\cdot}) = \lambda_0(T_{si}) exp\left[\beta_X^{SitePSIAdj} X_{si} + \boldsymbol{\beta}_S \boldsymbol{S}_{si} + \boldsymbol{\beta}_p^{1} \boldsymbol{e}_{1i}^{\cdot} S_{si}^{1} + \cdots + \boldsymbol{\beta}_p^{S} \boldsymbol{e}_{Si}^{\cdot} S_{si}^{S}\right], (5)$$

where $\boldsymbol{e}_{si}^{\cdot}$ is a (*K*-1)x1 vector of site-specific propensity score indicator variables for site *s*.

The second approach, Cox PH "Adj Site-PS B-Splines", uses b-spline basis from each site-specific propensity score model and adjusts for site and site-specific b-spline bases interacted with site. It is similar to model 5, except it replaces propensity score indicator variables with cubic b-spline bases as follows,

$$\lambda(T_{si}, \delta_{si} | X_{si}, \boldsymbol{S}_{si,} \boldsymbol{e}_{si}) = \lambda_0(T_{si}) exp\left[\begin{matrix} \beta_X^{SitePSBSAdj} X_{si} + \boldsymbol{\beta}_S \boldsymbol{S}_{si} + \boldsymbol{\beta}_p^{1} f(\boldsymbol{e}_{1i}^{\cdot}) S_{si}^{1} + \cdots \\ + \boldsymbol{\beta}_p^{S} f(\boldsymbol{e}_{Si}) S_{si}^{S} \end{matrix}\right], (6)$$

where $f(\boldsymbol{e}_{si})$ is a 5x1 vector of 5 cubic b-spline basis functions on the site-specific propensity score for site *s*.

D. SITE-STRATIFIED COX PH REGRESSION ADJUSTING FOR CATEGORICAL CONFOUNDERS OR SITE-SPECIFIC PROPENSITY SCORES (INDICATORS OR B-SPLINES)

Instead of adjusting for site and/or confounders in the mean model as outlined in model (1) in **Section II.A**, another common method to account for confounding by site and/or confounders is to use a stratified Cox PH regression model. The stratified cox model makes a proportional hazard assumption in each site and/or confounder strata but allows for different baseline hazards between strata.
A common approach is to do a site-stratified Cox PH model, but adjust for confounders directly in the regression (Stratify Site Adj Confounders). This approach accounts better for differences across sites compared with adjusting directly for site and therefore, reduces potential bias if there are different relationships between those that receive the exposure versus comparator across sites. The disadvantage of this approach is there may be some loss of power/efficiency relative to adjusting for site in the situation when a common baseline hazard assumption is valid. In the simulation study in **Section III**, we will assess whether this loss of power/efficiency occurs in the rare event setting. The specific form of the Cox PH regression model is,

$$\lambda(T_{si},\delta_{si}|X_{si},\boldsymbol{Z}_{si}) = \lambda_{s0}(T_{si})exp\left[\beta_x^{SiteStr}X_{si} + \boldsymbol{\beta}_z\ \boldsymbol{Z}_{si}\right], \text{for } s=1,\ldots,S. \quad (7)$$

This method still estimates a conditional HR ($\exp(\beta_x^{SiteStr})$), but is conditional on site as a strata and confounders as adjusted. However, if there are numerous confounders to adjust for one may have model fitting issues similar to problems when using the Cox PH adjusting for categorical confounders method outlined in **Section II.A**. We will also assess the performance of stratifying by site, but adjusting for site-specific propensity scores using cubic b-splines (Stratify Site Adj Site-PS B-splines) as follows,

$$\lambda\left(T_{si},\delta_{si}|X_{si},\boldsymbol{S}_{si,}\boldsymbol{e}_{si}\right) = \lambda_{s0}(T_{si})exp\left[\beta_X^{SiteStrPSBSAdj}X_{si} + \boldsymbol{\beta}_p^1 f(\boldsymbol{e}_{1i}^{\cdot})S_{si}^1 + \cdots + \boldsymbol{\beta}_p^S f(\boldsymbol{e}_{Si})S_{si}^S\right] \quad (8)$$

where $f(\boldsymbol{e}_{si})$ is a 5x1 vector of 5 cubic b-spline basis functions on the site-specific propensity score for site *s*. Another approach to dimension reduction would be to stratify on the propensity instead of adjust.

E. PROPENSITY SCORE-STRATIFIED COX PH REGRESSION

Stratifying on percentiles of propensity scores is another common approach to account for confounding. First, we will define the Cox PH "Stratify PS" method as stratifying on propensity score percentile strata. This method will include site as a confounder in the propensity score model similar to the methods outlined in **Sections II.B**. Then define the following stratified Cox PH model,

$$\lambda(T_{si},\delta_{si}|X_{si}) = \lambda_{k0}(T_{si})exp[\beta_x^{PSStr}X_{si}], \text{for } k=1,\ldots,K. \quad (9)$$

Previous literature indicated that 5 (*K*=5) quantiles of the propensity score was sufficient to account for confounding(7), but it depends on the strength of confounding and distribution of the propensity scores. Other literature indicated that 5 was actually not enough and residual confounding still persisted. (46) We will vary the number of quantiles to be 5, 10, 15, and 20 in the simulation evaluation. The potential advantage of this method relative to propensity score adjustment methods is the relaxation of the proportional hazard assumption between strata. However, often an overall propensity score model may not be viable to estimate in a distributed data setting, but site-specific propensity score models are estimable. The following section will outline such a method.

F. SITE AND SITE-SPECIFIC PROPENSITY SCORE-STRATIFIED COX PH REGRESSION

Another approach we will evaluate takes into account confounding by using stratified Cox PH regression in which the strata are defined as site and site-specific propensity score percentile strata. This Cox PH "Stratify Site+Site-PS" method has similar properties as the "Cox PH regression adjusting for site-specific propensity score indicators" method outlined in **Section II.C** except allowing the baseline proportional hazards function to vary by strata. The specific stratified Cox PH model fit is the following,

$$\lambda(T_{si}, \delta_{si}|X_{si}) = \lambda_{ks0}(T_{si}) exp\left[\beta_x^{SitePSStr} X_{si}\right], \text{for } k=1,\ldots,K \text{ and } s=1,\ldots,S. \quad (10)$$

This method relaxes the proportional hazards assumptions by assuming different baseline hazards across strata. However, this can lead to lowering power and modeling fitting issues as more strata are needed to control for confounding. Interacting strata with site will further increase the number of strata.

We will evaluate if this method will be viable in the simulation study for different Sentinel settings. We will compare methods in terms of bias, type I error, power, and coverage. Bias was defined as the difference between the estimated log HR and the true conditional log HR. Type I error was defined as the proportion of simulations that signaled (p-value<0.05 based on Score Test) when the true conditional log HR was set at 0. Power was defined as the proportion of simulations that signaled given the true conditional log HR was set at a specified value. Coverage was defined as the Wald 95% CI for the estimated HR included the true conditional HR.

Table 14. Summary of evaluated methods

Method	Confounder Control	Confounder Sharing
Adj Confounders+Site (1)	Regression on Confounders and Site	Pooled
Adj PS Indicators (3)	Regression on Propensity Score (includes confounders and site) Indicators	Pooled
Adj PS B-splines (4)	Regression on Propensity Score (includes confounders and site) B-Splines	Pooled
Stratify Site Adj Confounders (7)	Stratify on Site and regress on categorical confounders	Pooled
Stratify PS (9)	Stratify on Propensity Score (includes confounders and site) categories	Pooled
Adj Site-PS Indicators (5)	Regression on Site-Specific Propensity Scores (includes confounders only) Indicators and adjust for site and interactions with site	Site-Specific
Adj Site-PS B-splines (6)	Regression on Site-Specific Propensity Scores (includes confounders only) B-Splines and adjust for site and interactions with site	Site-Specific
Stratify Site Adj Site-PS B-splines (8)	Stratify on Site and regress on Site-Specific Propensity scores (includes confounders only) B-Splines and include interactions with site	Site-Specific
Stratify Site+Site-PS(10)	Stratify on Site and Site-Specific Propensity Scores (includes confounders only) categories	Site-Specific

IV. SIMULATION EVALUATION FOR THE NON-DISTRIBUTED DATA SETTING

The purpose of this simulation study was to compare the performance of regression and stratification methods using propensity scores (Non-Distributed Methods outlined in **Section III**) in a real world example where there are numerous sites with varying sample size, complex relationships between confounders (confounders are correlated), and complex relationships between confounders and exposure (e.g. the relationship between confounders and likelihood of receiving medication may be differential across sites). The reason we want to use actual data is to obtain realistic relationships between exposures of interest, confounders, and outcomes when assessing performance of methods. We will use the approach outlined in **Section II** to conduct the realistic data simulation. We will use two examples from the Mini-Sentinel Pilot study. We will first summarize the two studies in **Section IV.A** and then in **Section IV.B** we will provide results for the ACEI and Angioedema simulation study and **Section IV.C** we will provide results for the Rivaroxaban and Ischemic Stroke simulation study. In **Section V** and **VI** we will tailor and evaluate via simulation the most promising methods for the distributed data setting in which limited data is shared centrally by sites.

A. PREVIOUS STUDY SUMMARIES

1. ACEI and Angioedema Data Summary

Angiodema is an adverse effect that is known to be more common following use of angiotensin-converting enzyme inhibitors (ACEI) relative to other medications to control high blood pressure, such as beta blockers (BB).(11, 15, 22) A previous Mini-Sentinel task order(45) assessed the association between ACEI and angioedema -using a cohort from 2008 to 2012 within the Sentinel Network. This previous study used a new user cohort design in which participants were new users of either ACEI or BB and did not have a fill from either medication class in the 183 days before cohort entry. Once participants were eligible for the cohort, they did not allow re-entry in subsequent study years (only one exposure episode included per subject). After participants were enrolled in the study (index date) they were followed to determine time to first diagnosis of angioedema or censored due to disenrollment from healthcare plan, stopping use of the medication (+14 days added to follow-up time to allow for additional adverse events that may be related to medication use), or 364 days after the index date (interest in 1-year follow-up outcomes). A study summary is below, including the data available and cohort definitions.

Exposure of Interest: Angiotensin-converting enzyme inhibitor (ACEI)
Comparator: Beta Blocker (BB)
Outcome: Time to Angioedema or censoring

Sites: 5 Sentinel Sites

Confounders: Age (18-44, 45-54, 55-64, 65-99), Female(M/F), Charlston/Elixhauser Combined Comorbidity Score ([-2,0]/1+), Emergency Room Visits (0/1+), Inpatient Hospitalization (0/1+), Year drug initiated (2008, 2009, 2010, 2011, 2012)

Eligibility/Exclusion Criteria of Cohort: New user of ACEI or BB from 2008 to 2012; Continuous enrollment at their health plan with a drug benefit (defined as having a gap of less than 45 days between drug benefit enrollment periods) of at least 183 days prior to index date; Excluded if they had concomitantly used medications in both therapeutic classes of interest on the index date (i.e., filled more than 1 medication of interest on index); Excluded if they had a prior diagnosis of angioedema in the 183 days prior to the index date.

For this methods evaluation task order, we are reusing analytic datasets from 5 Sentinel Sites to conduct a simulation study to compare methods outlined in **Section III**, varying the relationship between the exposure of interest (ACEI) and the outcome (time to Angioedema or censoring).

2. Rivaroxaban and Ischemic Stroke Data Summary

We are using a subset of data assembled for the recently-conducted Mini-Sentinel Surveillance study, which evaluated if the new anticoagulant Rivaroxaban was associated with the adverse effect ischemic stroke compared to the comparator group Warfarin. The Mini-Sentinel surveillance study found an adjusted hazard ratio of 0.61 (95% CI: 0.47, 0.79) for the outcome Ischemic Stroke comparing Rivaroxaban new users to Warfarin new users. (43) The primary analysis used a 1 to M variable ratio propensity score exposure matching nested in each site to control for confounding with an outcome Cox PH model stratified by matched sets.

This report uses a subset of the original study data from two sites from 2013 to 2015 to conduct a simulation evaluation of the performance of the methods outlined in **Section III**. The study population is a new user cohort design in which participants were new users of either Rivaroxaban or Warfarin and did not have a fill from either medication class in the 183 days before cohort entry. We further restricted to those without a past history of cerebrovascular disease in the 183 days before cohort entry since this subset of the population has a lower outcome rate which is better to assess performance of the methods. Once participants were eligible for the cohort, re-entry in subsequent study years was not allowed (only one exposure episode included per subject). After participants were enrolled in the study (index date) they were followed to determine time to first diagnosis of ischemic stroke or censored due to disenrollment from the healthcare plan or stopping use of the medication (+7 days added to follow-up time to allow for additional adverse events that may be related to medication use). In this simulation, for simplicity, we further censored at 180 days after the index date since this is a new user cohort and follow-up time for most participants was less than 6 months. A study summary is below, including the data available and cohort definitions. Note we are using only a small subset of the >100 covariates used in the original study since most covariates were strongly correlated and several of them were measuring the same outcome (e.g. Peripheral Vascular Disease diagnostic codes and procedure codes were combined into a single confounder) to focus the simulation study. The subset of covariates was selected a priori based on knowledge about risk factors of the outcome. As will be shown, restricting to fewer covariates did not have much of an impact on the estimated risk in the study population.

Exposure of Interest: Rivaroxaban (RIVA)
Comparator: Warfarin (WARF)
Outcome: Time to ischemic stroke or censoring

Sites: 2 Sentinel Sites

Confounders: Age (21-55, 56-65, 66-75, and 76+); Sex (M/F); Charlston/Elixhauser Combined Comorbidity Score (-2-0, 1-4, and 5+); Emergency Room Visits (0/1+); Inpatient Hospitalization (0/1+); Year (2013, 2014, and 2015); Heart Failure/Cardiomyopathy (Y/N) , Hypertension (Y/N); Hyperlipidemia (Y/N); Coronary Artery Disease (Y/N with Yes including a code for Myocardial Infarction, Acute Coronary Syndrome, Percutaneous Coronary Intervention diagnostic or procedure, or Coronary Artery Bypass Graft diagnostic or procedure); Peripheral Vascular Disease (Y/N with Yes including Peripheral Vascular Disease diagnostic or procedure code and other Arterial Embolism); Diabetes (Y/N); Renal Disease (Y/N); and Tobacco (Y/N).

Eligibility/Exclusion Criteria of Cohort: Eligible patients were those with a new diagnosis of atrial fibrillation or atrial flutter who were a new user of RIVA or WARF after their AF diagnosis and from January 1, 2013 to April 30, 2015; and continuous enrollment at their health plan with a drug benefit (defined as having a gap of less than 45 days between drug benefit enrollment periods) of at least 183 days prior to index date. Patients were excluded if they had chronic dialysis, history of kidney transplant, end stage renal disease, mitral stenosis or mechanical heart valve, or recent joint replacement/arthroplasty surgery within 183 days before cohort entry. We further focused the analyses to the subgroup without a history of cerebrovascular disease in the 183 days before the index date (including any code for ischemic stroke, transient ischemic attack, other ischemic cerebrovascular disease diagnosis or procedure, and non-specific cerebrovascular symptoms) for the simulation study.

B. SIMULATION STUDY FOR ACEI AND ANGIOEDEMA EXAMPLE

1. ACEI and Angioedema Data Detailed

We are using the ACEI and Angioedema example previously described in **Section IV.A**. We will first briefly summarize the important aspects of the data we will be mimicking in our simulation study. **Table 15** shows the sample size and outcome information by site. The total sample size across all sites is 2,251,132 with smallest site (Site 5) having a total sample size of 62,857, while the largest site (Site 1) has a total sample size of 722,264. The average site sample size was 450,226. Therefore, the sample size at the sites is quite variable with two large sites (Site 1 and 4), two medium sites (Site 2 and 3), and one small site (Site 5).

The distribution of censoring times was driven primarily by the estimated time on drugs. In claims, the time on drugs is estimated from a stockpiling algorithm of days supply from consecutive filled prescriptions. Estimated time on drug had a multimodal distribution with distinct peaks (at 30 days, 60 days and 90 days) reflecting the typical days supply of 30 days for some prescriptions. At four of the five sites, the most common censoring time overall was 44 days (30 days + 14 day continuation period added to time on drug to allow for additional adverse events that may be related to medication use), which comprised approximately 25-30% of all censoring times at these four sites. At the remaining site the most common censoring time was 104 days which comprised approximately 37% of all censoring times. We will use the three most prevalent modes in the distribution of stockpiling of prescriptions times to help model our censoring distribution (common bumps are at 44, 104, and 194 days, but we will allow the data to choose the most common bumps which varied by site). Unadjusted outcome rates are relatively consistent across sites, yielding an unadjusted rate ratio of angioedema between 2.0 to 3.3 comparing ACEIs versus BBs.

Table 15. Sample size and outcome information by site and exposure group

Site	Exposure	N	Avg. Person-Days	Events	Events/ 1,000 P-years	Unadjusted Rate Ratios
Site 1	BB	315,378	107.0	236	2.55	2.0
	ACEI	406,886	128.9	728	5.07	
Site 2	BB	124,889	152.1	87	1.67	3.3
	ACEI	164,371	181.2	444	5.45	
Site 3	BB	209,281	129.9	130	1.75	3.0
	ACEI	291,955	154.1	647	5.25	
Site 4	BB	306,239	119.0	160	1.60	2.5
	ACEI	369,276	144.0	592	4.07	
Site 5	BB	28,942	114.7	23	2.53	2.0
	ACEI	33,915	149.9	70	5.03	

Further we present sample proportions for exposure and confounder levels by site in **Table 16**. We note that ACEIs were prescribed more often (~55%) than BBs (~45%) at all sites. Typically for a new medical product the exposure of interest would be less common than the comparator. We will explore such a new product example in the distributed portion of the report. For age, there are important differences across sites, and in particular, users of ACEI and BB are much older on average at Site 3 than at the other sites. Age is typically an important confounder (strong relationship to outcome) so methods using the confounders directly or site-specific propensity models may be preferable. Site 3 also has higher rates of comorbidities and inpatient visits which may be due to having an older population relative to other sites. There are also notable differences in the distribution of the year that drug exposure occurred; specifically, Site 1 had a much smaller proportion of drug exposures beginning in 2008 compared to the other sites.

Table 16. Exposure and confounder distributions by site

	Site 1	Site 2	Site 3	Site 4	Site 5
EXPOSURE					
BB	44	43	42	45	46
ACEI	56	57	58	55	54
CONFOUNDERS					
Age					
18-44	28	22	9	22	26
45-54	29	25	12	24	30
55-64	27	25	15	26	32
65+	17	28	63	28	13
Sex					
Male	51	49	48	51	52
Female	49	51	52	49	48
Comorbidity Score					
[-2,0]	79	78	64	76	77
1+	21	22	36	24	24
ED Visits					
0	81	81	84	87	79
1+	19	19	16	13	21
Inpatient Visits					
0	90	91	84	86	89
1+	10	9	16	14	11
Year					
2008	13	26	21	23	21
2009	27	23	21	22	21
2010	23	19	19	20	21
2011	19	16	19	18	19
2012	18	16	19	17	18

Data depicted is the column percent (%) showing the percent of each site's study population with a given exposure or confounder.

Table 17 shows the relationship between the confounders and the exposure of interest by site. These are the coefficients from the observed data's site-specific propensity score models estimating the propensity of being exposed to the ACEI relative to BB given the confounder conditional on all other confounders. There seems a similar propensity of being given ACEI relative to BB across all sites for age (ACEI most likely amongst those 45-64, medium likely 65+ and least likely 18-44). Site 5 shows ACEI being given at an even higher likelihood across all older age groups relative to other sites. ACEI are less likely to be given to Females compared to Males, but this relationship is less strong amongst those at Site 3. Those who receive ACEI are less likely to have any comorbidities, ED Visits, or Inpatient Visits and these relationships seem to be consistent across sites. There are some modest site differences in uptake of ACEI over the study years in which some sites had higher propensity to prescribe ACEI relative to BB starting in 2009 (Site 3, Site 4, and Site5 indicated by higher odds ratios) while the other two sites had very similar propensity to prescribe ACEI across all study years.

Table 17. Odds ratios for confounders regressed on exposure (ACEI) by site (propensity score models)

	Site 1	Site 2	Site 3	Site 4	Site 5
Age (Ref: 18-44)					
45-54	1.68	1.76	1.55	1.60	1.84
55-64	1.64	1.73	1.48	1.56	1.84
65+	1.28	1.31	1.29	1.32	1.53
Sex (Ref: Male)	0.61	0.62	0.85	0.64	0.58
Comorbidity Score 1+	0.55	0.60	0.65	0.54	0.55
1+ ED Visits	0.82	0.72	0.84	0.85	0.80
1+ Inpatient Visits	0.52	0.49	0.49	0.50	0.41
Year (Ref: 2008)					
2009	1.08	1.02	1.11	1.11	1.11
2010	1.07	1.05	1.13	1.12	1.14
2011	1.04	0.96	1.10	1.07	1.13
2012	1.01	0.91	1.09	1.02	1.13

Note that a single model is run within each site and therefore each odds ratio is conditional on all other confounders.

Figure 8. Histogram showing the overlap of the propensity score distributions by exposure and site

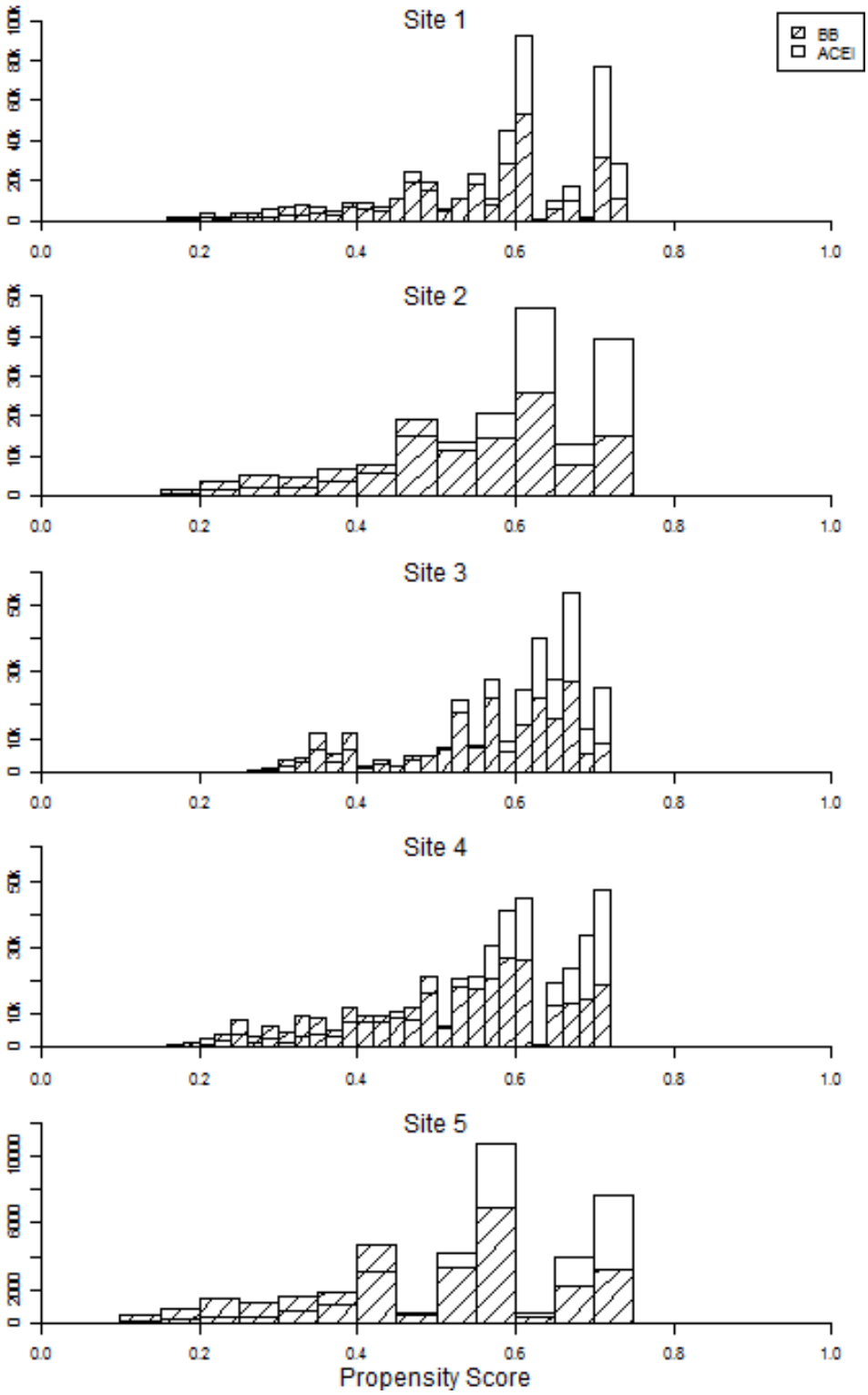


Figure 8 shows the overlap between propensity score across exposure groups by site. There is very good overlap indicating that the assumption of positivity is likely met in this population (e.g. everyone has potential to receive either drug in the population) given the covariates we had available.

Table 18 a and Table 18 b show the adjusted hazard ratios fitting site-specific survival models including the exposure of interest (ACEI) and all confounders in each model. **Table 18 a** shows the results fitting a Cox Proportional Hazards Model (Cox PH) while **Table 18 b** shows the results fitting a site-specific Weibull Accelerated Failure time model. Note that both models assume proportional hazards, but the Weibull Accelerated Failure time model assumes a flexible Weibull distribution on the outcome time to The tables indicate that both models estimate extremely similar hazards ratios and therefore the data is not sensitive to adding the additional assumption of the Weibull distribution.

The findings from this example show that ACEI has a higher rate of angioedema relative to BB and that adjusted hazard ratios range between 2.4 and 3.6 across sites. However, the relationship between the confounders and angioedema is not consistent across sites for age, sex, ED visits, or year. Therefore, there is potential for differential relationships between outcome and confounders by site.

We further show in **Figure 9** that both the adjusted HR and estimated 95% comparing ACEI to BB are extremely similar across all sites between these two models. This finding is important, as the data generation program simulates data assuming the Weibull Accelerated Failure time model while all of the methods we will be evaluating in the simulation study assume a Cox PH model framework.

Table 18 a. Adjusted hazard ratios for exposure of interest (ACEI) and confounders from site-specific cox proportional hazards models

	Site 1	Site 2	Site 3	Site 4	Site 5
EXPOSURE					
ACEI (Ref: BB)	2.41	3.64	3.40	2.98	2.39
CONFOUNDERS					
Age (Ref: 18-44)					
45-54	0.97	1.27	0.86	1.19	1.75
55-64	0.90	1.27	0.99	0.93	1.45
65+	0.89	1.14	1.08	1.19	1.43
Sex (Ref: Male)	1.20	1.08	1.09	0.99	1.59
Comorbidity Score 1+	1.50	1.50	1.38	1.54	1.86
1+ ED Visits	0.89	0.78	1.18	1.24	0.82
1+ Inpatient Visits	1.93	1.34	1.20	1.59	1.36
Year (Ref: 2008)					
2009	1.11	1.08	1.00	0.94	1.33
2010	1.17	0.88	1.02	1.17	1.64
2011	1.07	0.93	1.28	1.06	0.88
2012	1.20	0.88	1.01	0.93	1.49

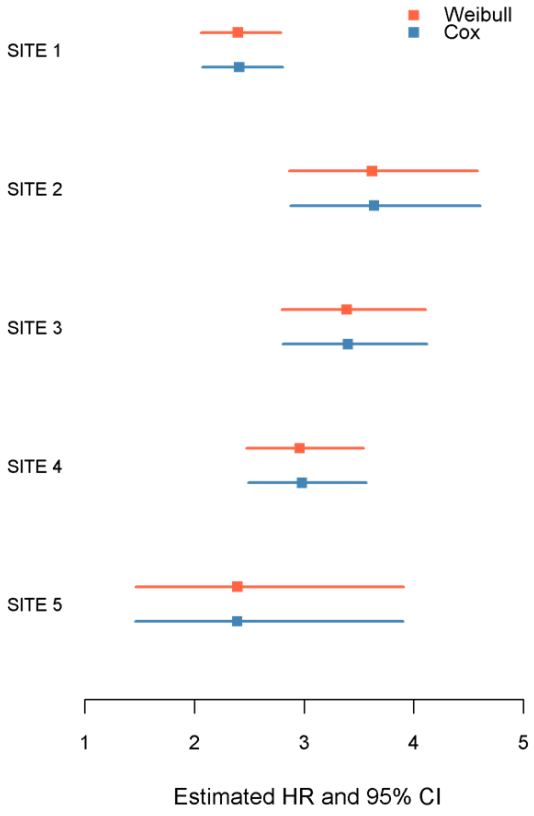
Note that a single model is run within each site and therefore each hazard ratio is conditional on all other covariates in the model.

Table 18 b. Adjusted hazard ratios for exposure of interest (ACEI) and confounders from site-specific Weibull accelerated failure time models

	Site 1	Site 2	Site 3	Site 4	Site 5
EXPOSURE					
ACEI (Ref: BB)	2.40	3.62	3.39	2.96	2.39
CONFOUNDERS					
Age (Ref: 18-44)					
45-54	0.97	1.27	0.85	1.19	1.76
55-64	0.90	1.28	0.99	0.93	1.47
65+	0.89	1.15	1.07	1.17	1.45
Sex (Ref: Male)	1.20	1.08	1.09	1.00	1.58
Comorbidity Score 1+	1.51	1.50	1.38	1.55	1.85
1+ ED Visits	0.89	0.79	1.19	1.25	0.82
1+ Inpatient Visits	1.93	1.34	1.20	1.59	1.36
Year (Ref: 2008)					
2009	1.11	1.08	1.00	0.94	1.33
2010	1.17	0.88	1.02	1.17	1.65
2011	1.07	0.93	1.28	1.06	0.87
2012	1.19	0.88	1.01	0.92	1.44

Note that a single model is run within each site and therefore each hazard ratio is conditional on all other covariates in the model

Figure 9. Hazard ratios and 95% CIs by site for Weibull and Cox time-to-event models



2. Simulation Generation and Evaluation Study

To mimic this ACEI and Angioedema real world data example, we generated simulated data using the framework detailed in **Section II**, but we will briefly summarize here. For each site we calculated the following summary statistics:

- Confounders: Probabilities within Confounders (**Table 16**) and Common Probabilities between each confounder category and all other confounders
- Exposure|Confounders (Propensity Model): Coefficients from a logistic model fitting the outcome ACEI versus BB with all covariates in the model (**Table 17**)
- Outcome|Exposure and Confounders: Coefficients from a Weibull Accelerated Failure time for the outcome time to angioedema including exposure and confounders in the model. We use the true coefficients for the confounder variables at each site (**Table 18 b**) and alter the exposure coefficients depending on the strength of relationship desired.
- Censored|Exposure, Confounders: Allowed for the three most prevalent modes in prescribing patterns (typically 30, 90, and 180 days, but we allowed the data to choose the most common modes, so they varied by site) and returned the prevalence of the modes and time of each mode (See **Table 8** for common censoring modes by site). To model the censoring distribution amongst those that were not censored at any of the three most common prescribing modes we obtained coefficients from the Weibull Accelerated Failure time model for the outcome time to censoring (censor now the outcome) (**Table 10**) and we censored in this model at the time of angioedema (angioedema now the censor variable). This model was fit amongst only those that were not censored at the three most common prescribing modes and did not include covariates.

These summary statistics were then used to simulate subject level datasets independently for each site, including simulated covariates, exposure and time-to-event or censoring. See **Section II** for details of how to simulate such data.

We performed 2,000 simulations for each of the four different treatment effect scenarios using total sample sizes of 150,000 distributed by the proportionate size of each site in the example datasets. In the first three scenarios, the HR comparing ACEI with BB was set to 1.0, 1.5 and 2.0 and was the same at each site (homogeneous). The fourth scenario allowed the HR to be heterogeneous/vary by site in the same way that the estimates varied in the observed example data (**Table 18 b** and **Figure 9**). To calculate the pooled HR estimate in the setting where the HR is heterogeneous we fit a site stratified Cox PH model with all confounders and a single term for the effect of exposure (Model 7). This estimate of the HR was used as the truth for the simulations that included heterogeneity of the effect of exposure on outcome. For each set of simulated data, all proposed models were fit and the resulting estimates, standard errors, test statistics and hypothesis tests returned. Estimates of bias (on the log HR scale), power (using log rank tests) and coverage (log HR scale using Wald Confidence intervals) are presented in **Table 19** and **Table 20**.

3. ACEI and Angioedema Simulation Study Results

Simulation results are presented in **Table 19** and **Table 20**. Overall the results were favorable for all estimators. When the treatment effect is homogeneous among the sites (**Table 19**), pooled analysis methods – adjusting directly for confounders, adjusting for deciles of the propensity score or adjusting for the propensity score using B-splines – and stratification methods – stratifying on site and adjusting for confounders, or stratifying on deciles of the propensity score – were found to have comparable performance in terms of bias, type I error, power, and coverage. However, a key finding was that using

quintiles of the propensity score may provide insufficient control of confounding, whether used for adjustment or stratification. Therefore, caution should be taken in the Sentinel context when categorizing the propensity score. Partitioning into at least deciles whenever possible is recommended, especially when using a pooled propensity score model that assumes a consistent relationship between confounders and exposure conditional on site. Models with site-specific propensity scores tended to have smaller bias and higher power than their pooled-data counterparts. Nominal coverage was achieved by all estimators.

In the setting with small amounts of site heterogeneity (**Table 20**) in which site-specific models should theoretically outperform pooled data models, we did not find an appreciable difference. This likely reflects the moderate differences observed across sites: HRs of 2.40, 3.62, 3.39, 2.96, and 2.39, respectively, and is a limitation of the example we are using for this final report.

Note at the bottom of each of the tables we present a series of estimates for reference, including a marginal estimator and the unadjusted estimator. Our marginal estimator estimates a HR for the average treatment effect (ATE) for the entire population by first estimating a conditional Cox PH regression model and then marginalizing that estimate to the ATE population by estimating the HR assuming everyone was treated compared to everyone remaining untreated similar to Austin 2013(40). When we state we are estimating a Marginal Stratified model, our underlying conditional Cox PH model adjusts for site-specific covariates and site is used as a stratification variable (Model 7). For the marginal simulated value, we use simulated datasets and calculate the marginal estimate on each dataset and present the mean estimate. We further show the unadjusted estimates to gauge the magnitude of the bias introduced by confounders. Across all scenarios the observed relative bias was approximately 10% in the direction of the null. Since we are in the rare outcome setting, the conditional HR and the marginal HR are approximately identical due to collapsibility conditions. Therefore, when we show bias we are assessing the very small difference between the marginal and conditional HR for this setting.

Table 19. Simulation results with homogeneous effects across sites (5 sites, 2,000 simulations, samples of size 150,000)

	HR = 1.0			HR = 1.5			HR = 2.0		
Model	**Bias**	**Type I**	**Coverage**	**Bias**	**Power**	**Coverage**	**Bias**	**Power**	**Coverage**
Pooled Data									
Adj Confounders + Site	0.003	0.038	0.961	0.0003	0.670	0.956	0.006	0.989	0.959
Adj PS Indicators									
5 quantiles	-0.010	0.035	0.958	-0.013	0.648	0.953	-0.008	0.987	0.956
10 quantiles	-0.003	0.039	0.962	-0.005	0.657	0.954	0.000	0.989	0.957
15 quantiles	-0.003	0.035	0.963	-0.005	0.658	0.955	0.000	0.987	0.958
20 quantiles	0.000	0.037	0.962	-0.002	0.664	0.955	0.003	0.987	0.957
Adj PS B-splines	0.002	0.038	0.961	-0.001	0.664	0.957	0.005	0.989	0.958
Stratify Site Adj Conf	0.003	0.038	0.962	0.000	0.670	0.955	0.006	0.989	0.959
Stratify PS									
5 quantiles	-0.010	0.039	0.958	-0.013	0.666	0.953	-0.008	0.989	0.956
10 quantiles	-0.003	0.041	0.962	-0.005	0.678	0.954	0.000	0.989	0.957
15 quantiles	-0.003	0.039	0.963	-0.005	0.677	0.954	0.000	0.990	0.957
20 quantiles	0.000	0.042	0.962	-0.002	0.681	0.957	0.003	0.991	0.957
Site-Specific									
Adj Site-PS Indicators									
5 quantiles	-0.004	0.034	0.958	-0.008	0.659	0.953	-0.003	0.988	0.956
10 quantiles	0.000	0.038	0.959	-0.002	0.662	0.955	0.003	0.988	0.957
15 quantiles	0.004	0.039	0.961	0.001	0.668	0.955	0.007	0.989	0.958
20 quantiles	0.004	0.040	0.958	0.002	0.664	0.954	0.007	0.989	0.956
Adj Site-PS B-splines	0.006	0.044	0.959	0.004	0.670	0.956	0.009	0.987	0.958
Stratify Site+Site-PS									
5 quantiles	-0.005	0.038	0.957	-0.008	0.677	0.951	-0.003	0.989	0.956
10 quantiles	0.000	0.044	0.959	-0.002	0.677	0.953	0.003	0.990	0.957
15 quantiles	0.004	0.045	0.961	0.001	0.685	0.955	0.007	0.989	0.956
20 quantiles	0.004	0.045	0.959	0.002	0.687	0.953	0.007	0.990	0.955
Stratify Site Adj B-splines	0.006	0.035	0.961	0.002	0.666	0.954	0.006	0.988	0.955
Reference Estimators Not for Methods Comparison									
Marginal Simulated	0.006		0.955	0.005		0.952		0.945	0.006
Unadjusted	-0.092		0.933	-0.095		0.916		0.914	-0.092

* Follow the hyperlinks to find detailed descriptions of each method.

Table 20. Simulation results with observed treatment heterogeneity across sites (5 sites, 2000 simulations)

	HR = 2.94*		
Model	**Bias**	**Power**	**Coverage**
Pooled Data			
Adj Confounders + Site	-0.009	1.000	0.952
Adj PS Indicators			
5 quantiles	-0.022	1.000	0.947
10 quantiles	-0.014	1.000	0.947
15 quantiles	-0.013	1.000	0.949
20 quantiles	-0.010	1.000	0.951
Adj PS B-splines	-0.009	1.000	0.950
Stratify Site Adj Conf	-0.009	1.000	0.951
Stratify PS			
5 quantiles	-0.022	1.000	0.947
10 quantiles	-0.014	1.000	0.947
15 quantiles	-0.013	1.000	0.949
20 quantiles	-0.010	1.000	0.950
Site-Specific			
Adj Site-PS Indicators			
5 quantiles	-0.019	1.000	0.948
10 quantiles	-0.012	1.000	0.949
15 quantiles	-0.008	1.000	0.951
20 quantiles	-0.008	1.000	0.953
Adj Site-PS B-splines	-0.006	1.000	0.953
Stratify Site+Site-PS			
5 quantiles	-0.019	1.000	0.948
10 quantiles	-0.013	1.000	0.950
15 quantiles	-0.008	1.000	0.951
20 quantiles	-0.008	1.000	0.953
Stratify Site Adj B-splines	-0.011	1.000	0.949
Reference Estimators Not for Methods Comparison			
Marginal Stratified	0.000		
Unadjusted	-0.100	0.000	0.894

* HR = 2.94 is the observed stratified estimate from the actual example.

C. SIMULATION STUDY FOR RIVAROXABAN AND ISCHEMIC STROKE EXAMPLE

1. Rivaroxaban and Ischemic Stroke Data Detailed

We are using the rivaroxaban and ischemic stroke example previously described in **Section IV.B**. We will first briefly summarize the important aspects of the data which we will be mimicking in our simulation study. **Table 21** shows the sample size and outcome information by site for the entire cohort, the subgroup with no prior history of ischemic stroke in the last 183 days, and the subgroup with no prior history of cerebrovascular disease in the last 183 days. We will use the last subgroup for the simulation study since for statistical purposes having a rarer outcome with events not bunched so strongly at the beginning of the follow-up time will test methods better, but we show the results for all groups since it may be of interest to see the change in population and effect overall and within the subgroups. The total sample size across both sites is 39,197 with 15,972 (40.7%) exposed to Rivaroxaban (RIVA). A total of 659 ischemic stroke events were observed with 471 amongst WARF users and 188 amongst RIVA users. Unadjusted rate ratios showed a protective effect for using RIVA relative to WARF (overall RR=0.610). When focusing on the primary subgroup with no prior cerebrovascular disease, the total sample size was 30,502 (78% of the entire study population) with 12,830 (42.1%) exposed to RIVA. Within this subgroup outcome rates drop substantially from 40.1 events per 1,000 people over 180 days to 17.1 amongst WARF users and 24.5 events per 1,000 people over 180 days to 14.0 amongst RIVA users. The unadjusted rate ratio was also attenuated to 0.816 overall.

Table 21. Sample size and outcome information by site and exposure group

Site	Exposure	N	Avg. Person-Days	Events	Events/1,000 P-180 days	Unadjusted Rate Ratios
Entire Cohort						
Overall	WARF	23225	91.0	471	40.1	0.610
	RIVA	15972	86.6	188	24.5	
Site 1	WARF	18258	94.3	393	41.1	0.593
	RIVA	13298	89.3	161	24.4	
Site 2	WARF	4967	78.8	78	35.9	0.697
	RIVA	2674	72.8	27	25.0	
No Prior Ischemic Stroke						
Overall	WARF	20492	91.5	216	20.7	0.729
	RIVA	14586	86.5	106	15.1	
Site 1	WARF	16084	94.9	183	21.6	0.708
	RIVA	12121	89.4	92	15.3	
Site 2	WARF	4408	78.9	33	17.1	0.829
	RIVA	2465	72.3	14	14.1	
No Prior Cerebrovascular Disease						
Overall	WARF	17672	91.7	154	17.1	0.816
	RIVA	12830	86.5	86	14.0	
Site 1	WARF	13779	95.3	126	17.3	0.813
	RIVA	10593	89.5	74	14.0	
Site 2	WARF	3893	79.0	28	16.4	0.819
	RIVA	2237	71.9	12	13.4	

As shown in **Figure 10**, in the entire population, the most common censoring time was 44 days (30 days +14-day continuation period) overall for both medications and sites and had modes at other common prescription durations of 60 days, 90 days and so on as well as 180 days (end of follow-up period). Further as shown in **Figure 11** most ischemic stroke outcomes occurred within the first 15 to 30 days. In contrast to the ACE inhibitor example where the outcome, angioedema, is an allergic reaction and often occurs soon after drug initiation, ischemic strokes would not be expected to occur rapidly following exposure to an anticoagulant. **Figure 12** shows that amongst those with no prior history of cerebrovascular disease, censoring was still most common at 44 days (30 days +14-day continuation period) and had bumps at other common prescription durations of 60 days, 90, and so on as well as 180 days (end of follow-up period). Further, **Figure 13** shows ischemic stroke outcomes are more evenly distributed over time than in the overall population, but with slight elevation in the first 30 days of the study period. This pattern is likely partly due to the fact that most of the person-time available is early in the observation period.

Figure 10. Histogram of time to censoring by site and exposure group in the entire cohort (n=39,197 at both sites combined)

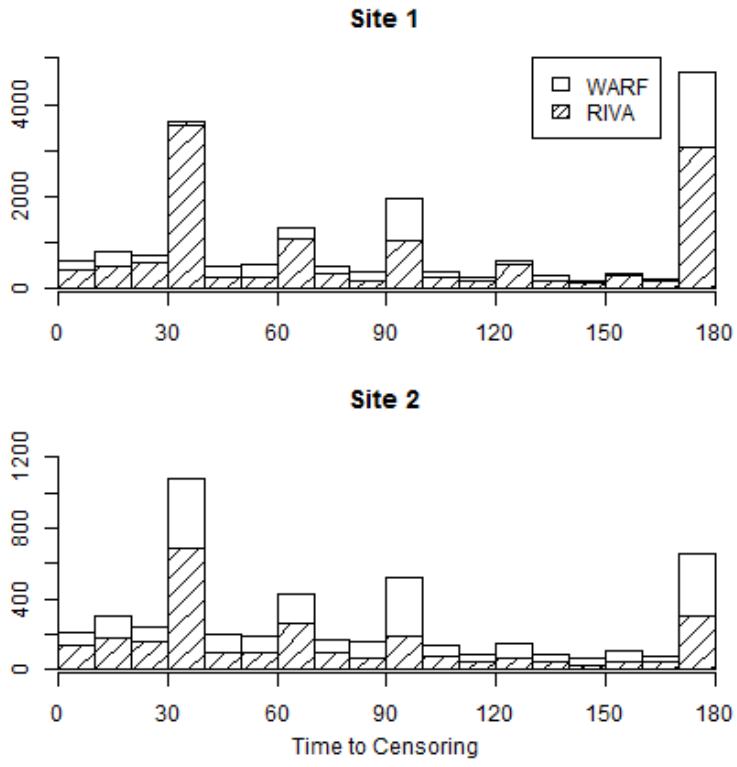


Figure 11. Histogram of time to ischemic stroke by site and exposure group (entire cohort)

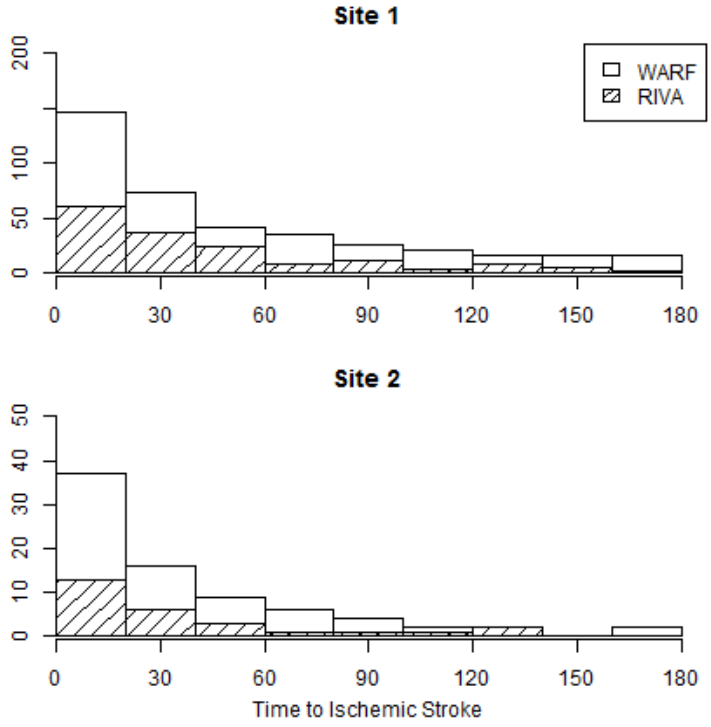


Figure 12. Histogram of time to censoring by site and exposure group amongst those without history of cerebrovascular disease (n=30,502 at both sites combined)

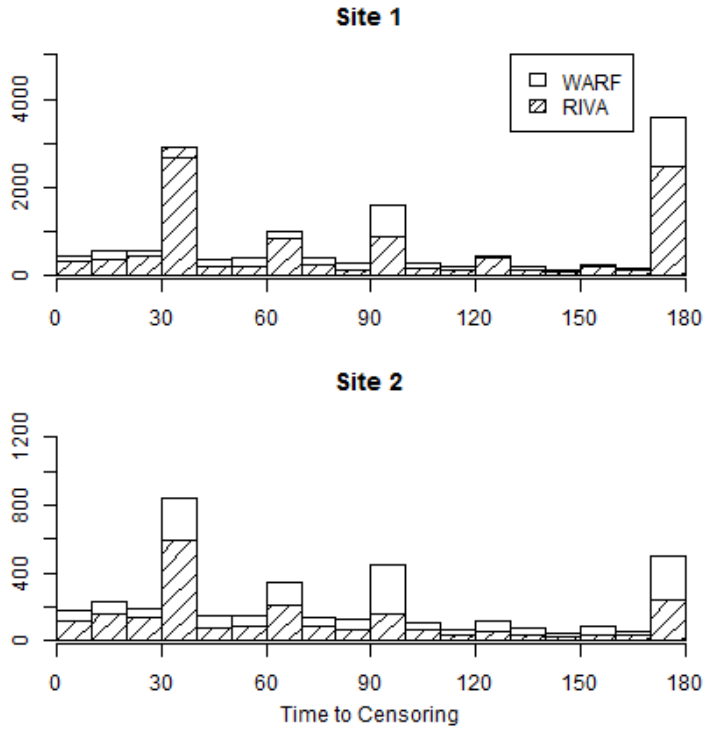


Figure 13. Histogram of time to ischemic stroke by site and exposure group among those without history of cerebrovascular disease

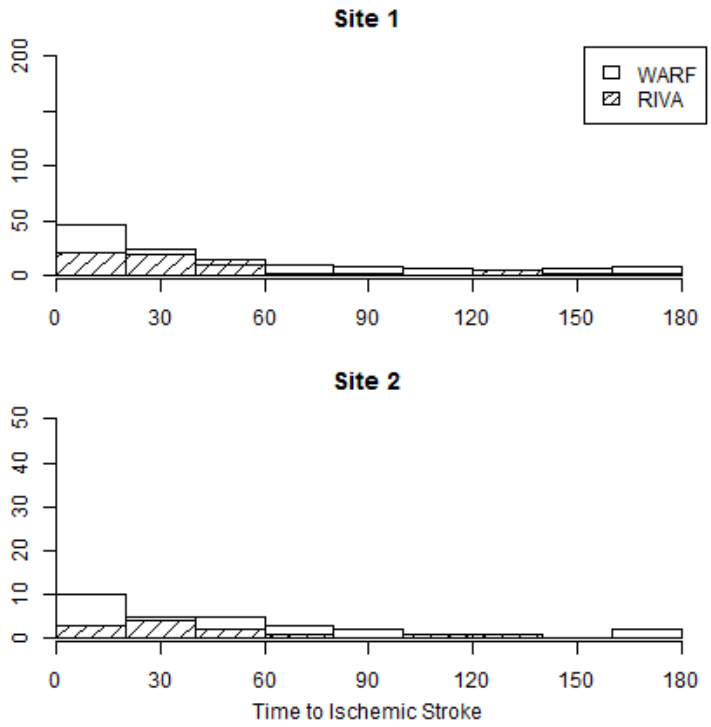


Table 24 presents sample proportions for exposure and confounder levels by site and different cerebrovascular disease subgroups. We note that WARF was prescribed more often than RIVA at both sites. For age, there are important differences by sites, and in particular, users of medications were older at Site 1 relative to Site 2. This age differences likely yields the observed lower comorbidity index score in Site 2 relative to Site 1. In general rates of most cardiovascular and renal outcomes are lower in Site 2 than in Site 1 indicating a healthier user population. Note that we do not have data for 2015 for Site 2. As we will show Year is not strongly related to outcome and therefore for methods purposes we removed it as a potential confounder in our simulation framework.

Table 22. Exposure and confounder distributions by site and cerebrovascular disease subgroups

	Everyone		No Prior Ischemic Stroke		No Prior Cerebrovascular Disease	
	Site1	Site2	Site1	Site2	Site1	Site2
EXPOSURE						
WARF	57.9	65.0	57.0	64.1	56.5	63.5
RIVA	42.1	35.0	43.0	35.9	43.5	36.5
CONFOUNDERS						
Age						
21-55	2.0	6.7	2.0	7.0	2.2	7.7
56-65	7.8	16.5	7.9	17.4	8.2	18.5
66-75	38.3	27.6	38.9	27.8	39.4	27.7
76+	52.0	49.2	51.2	47.7	50.2	46.1
Sex (Female)	45.8	42.2	45.1	40.8	44.9	40.5
Comorbidity Score						
-2-0	15.8	21.3	16.9	22.5	18.2	23.8
1-4	53.0	59.0	53.8	59.4	54.6	59.4
≥5	31.2	19.7	29.3	18.0	27.2	16.8
ER Visits (1+)	29.2	19.6	27.7	19.2	26.3	18.3
Inpatient Visits (1+)	52.0	51.6	48.5	47.8	46.0	45.8
Year						
2013	24.1	63.0	24.1	62.9	24.0	62.8
2014	56.7	37.0	56.6	37.1	56.6	37.2
2015	19.2	0.0	19.2	0.0	19.3	0.0
Cerebrovascular Disease	22.8	19.8	13.6	10.8	0.0	0.0
Heart Failure/Cardiomyopathy	46.7	38.8	45.7	37.6	44.5	36.4
Hypertension	86.5	77.9	85.6	76.6	84.6	75.2
Hyperlipidemia	30.1	27.6	29.8	26.6	28.8	25.5
Coronary Artery Disease	27.1	19.9	26.5	19.0	24.1	17.3
Peripheral vascular disease	25.4	17.5	24.4	16.4	21.6	14.1
Diabetes	39.7	29.9	39.2	29.0	38.3	28.1
Renal Disease	24.9	14.1	24.3	13.4	23.3	12.6
Tobacco	19.5	10.6	19.2	10.4	18.5	10.3

*Data depicted is the column percent (%) showing the percent of each site's study population with a given exposure or confounder.

Table 23. Odds ratios for confounders regressed on exposure (RIVA) by site (propensity score models)

	Everyone		No Prior Ischemic Stroke		No Prior Cerebrovascular Disease	
	Site1	Site2	Site1	Site2	Site1	Site2
Age (Ref:21-55)						
56-65	0.86	0.80	0.85	0.77	0.80	0.77
66-75	0.86	0.45	0.84	0.43	0.81	0.43
76+	0.71	0.33	0.68	0.30	0.63	0.29
Sex (Ref: Male)	1.07	1.02	1.06	1.05	1.05	1.08
Comorbidity Score (Ref:-2-0)						
1-4	0.81	0.87	0.81	0.87	0.82	0.85
≥5	0.57	0.55	0.57	0.53	0.58	0.54
1+ ER Visits	1.12	1.21	1.12	1.21	1.12	1.26
1+ Inpatient Visits	1.17	1.09	1.22	1.11	1.26	1.11
Year (Ref: 2013)						
2014	1.25	1.66	1.26	1.68	1.24	1.69
2015	1.23		1.24		1.26	
Cerebrovascular Disease	0.83	0.90	0.93	1.00		
Heart Failure/Cardiomyopathy	0.89	0.94	0.88	0.95	0.88	0.94
Hypertension	1.04	0.99	1.04	0.97	1.02	0.97
Hyperlipidemia	1.13	1.36	1.14	1.35	1.17	1.32
Coronary Artery Disease	0.95	0.76	0.94	0.76	0.94	0.80
Peripheral vascular disease	0.95	0.97	0.94	1.00	0.94	0.96
Diabetes	0.87	0.79	0.86	0.83	0.84	0.85
Renal Disease	0.92	0.80	0.91	0.82	0.90	0.76
Tobacco	1.18	1.02	1.18	1.06	1.18	1.01

*Data depicted is the column percent (%) showing the percent of each site's study population with a given exposure or confounder.

Table 26 shows the relationship between the confounders and the exposure of interest by site and different cerebrovascular disease subgroups. These are the coefficients from the observed data's site-specific propensity score models estimating the propensity of being exposed to RIVA relative to WARF given the confounder conditional on all other confounders. There is a lower propensity to give RIVA in general to older age groups, but in particular in Site 2 relative to Site 1 indicating that age is likely a very strong confounder that is differential across site. Those with a higher comorbidity index are also more likely to be given WARF then RIVA which is similar across sites. Those with worse renal function are also more likely to be given WARF relative to RIVA. Both high comorbidity index and poor renal function are also related to outcome. Therefore, there are several potentially strong confounders.

Table 24. Odds ratios for confounders regressed on exposure (RIVA) by site (propensity score models)

	Everyone		No Prior Ischemic Stroke		No Prior Cerebrovascular Disease	
	Site1	Site2	Site1	Site2	Site1	Site2
Age (Ref:21-55)						
56-65	0.86	0.80	0.85	0.77	0.80	0.77
66-75	0.86	0.45	0.84	0.43	0.81	0.43
76+	0.71	0.33	0.68	0.30	0.63	0.29
Sex (Ref: Male)	1.07	1.02	1.06	1.05	1.05	1.08
Comorbidity Score (Ref:-2-0)						
1-4	0.81	0.87	0.81	0.87	0.82	0.85
≥5	0.57	0.55	0.57	0.53	0.58	0.54
1+ ER Visits	1.12	1.21	1.12	1.21	1.12	1.26
1+ Inpatient Visits	1.17	1.09	1.22	1.11	1.26	1.11
Year (Ref: 2013)						
2014	1.25	1.66	1.26	1.68	1.24	1.69
2015	1.23		1.24		1.26	
Cerebrovascular Disease	0.83	0.90	0.93	1.00		
Heart Failure/Cardiomyopathy	0.89	0.94	0.88	0.95	0.88	0.94
Hypertension	1.04	0.99	1.04	0.97	1.02	0.97
Hyperlipidemia	1.13	1.36	1.14	1.35	1.17	1.32
Coronary Artery Disease	0.95	0.76	0.94	0.76	0.94	0.80
Peripheral vascular disease	0.95	0.97	0.94	1.00	0.94	0.96
Diabetes	0.87	0.79	0.86	0.83	0.84	0.85
Renal Disease	0.92	0.80	0.91	0.82	0.90	0.76
Tobacco	1.18	1.02	1.18	1.06	1.18	1.01

Figure 10 shows the overlap between propensity score across exposure groups by site amongst those without prior cerebrovascular disease (figures look similar for everyone and those with prior ischemic stroke; data not shown). There is very good overlap indicating that the assumption of positivity is likely met in this population (e.g. everyone has potential to receive either drug in the population) given the covariates we had available.

Figure 14. Histogram showing the overlap of the propensity score distributions by exposure and site amongst those without history of cerebrovascular disease

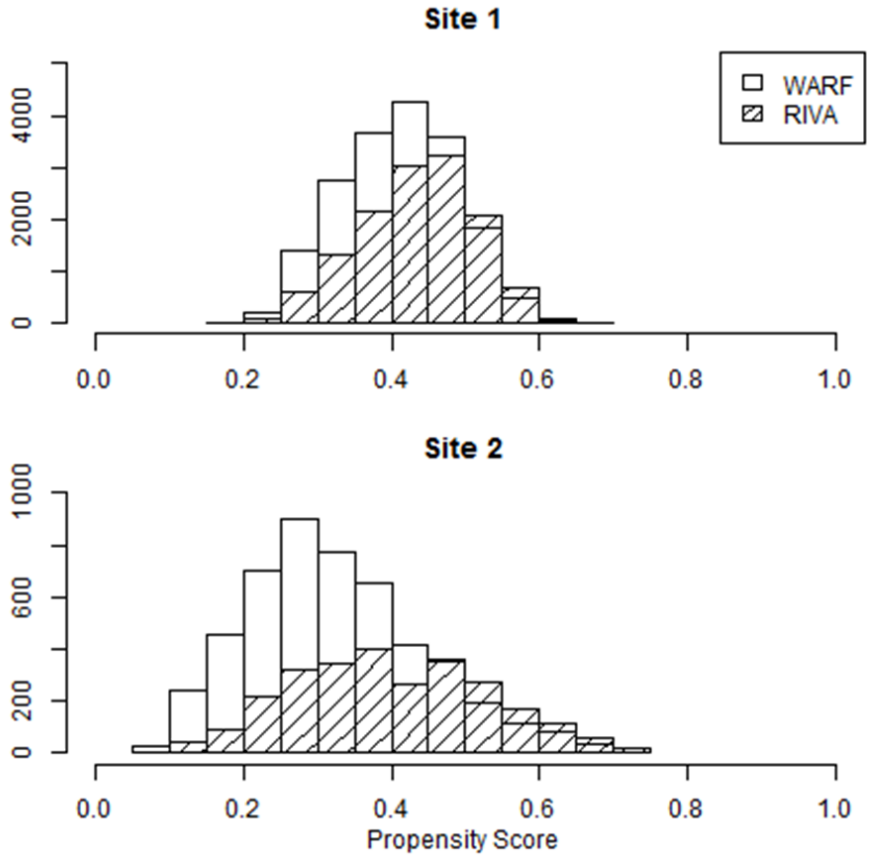


Table 25 a and Table 25 b show the adjusted hazard ratios fitting site-specific survival models including the exposure of interest (RIVA) and all confounders in each model. **Table 25 a** shows the results fitting a Cox Proportional Hazards Model (Cox PH) while **Table 25 b** shows the results fitting a site-specific Weibull Accelerated Failure time model.

The findings from this analysis show that in the entire population RIVA has an overall adjusted hazard ratio of 0.70 (0.58, 0.83) of ischemic stroke relative to WARF and that the adjusted hazard ratios range from 0.67 to 0.89 across sites. The overall adjusted hazard ratio was calculated pooling the sites data and running a Cox PH model adjusting for all confounders and site in a single model. However, when we restrict the analyses to the subset to ~80% of the population with no prior cerebrovascular disease, the adjusted hazard ratio is attenuated to 0.90 (0.68, 1.17) ranging from 0.88 to 0.94 across the sites. Further, the relationship between the confounders, exposure, and ischemic stroke are not consistent across sites for age, comorbidity index, cardiovascular disease, and renal disease. Therefore, there is potential for differential relationships between outcome and confounders by site.

Table 25 a. Adjusted hazard ratios for ischemic stroke by exposure of interest (RIVA) and confounders from site-specific Cox proportional hazards models

	Everyone		No Prior Ischemic Stroke		No Prior Cerebrovascular Disease	
	Site1	Site2	Site1	Site2	Site1	Site2
EXPOSURE						
RIVA	0.67	0.89	0.78	0.97	0.88	0.94
CONFOUNDERS						
Age (Ref:21-55)						
56-65	0.91	2.26	0.54	2.63	0.55	2.55
66-75	0.84	1.59	0.66	1.67	0.83	1.31
76+	1.12	2.48	1.07	2.55	1.14	2.08
Sex (Ref: Male)	1.26	1.21	1.39	1.18	1.37	1.37
Comorbidity Score (Ref:-2-0)						
1-4	1.05	1.26	1.08	1.06	1.14	0.94
≥5	1.53	2.32	1.78	1.80	1.95	1.72
1+ ER Visits	1.00	1.05	0.90	1.22	0.84	1.12
1+ Inpatient Visits	2.06	1.77	1.52	0.93	1.37	0.90
Year (Ref: 2013)						
2014	1.05	0.98	1.09	1.00	0.98	1.15
2015	1.00		1.05		0.82	
Cerebrovascular Disease	4.70	4.96	1.98	1.09		
Heart Failure/Cardiomyopathy	1.06	1.25	1.02	1.04	1.18	1.04
Hypertension	0.82	1.45	0.69	1.64	0.63	1.90
Hyperlipidemia	1.21	0.82	1.14	0.71	1.07	0.61
Coronary Artery Disease	0.78	1.10	0.87	1.63	1.02	1.71
Peripheral vascular disease	0.99	0.58	1.08	1.22	0.95	1.73
Diabetes	1.10	1.10	1.21	1.12	1.20	1.04
Renal Disease	1.01	0.68	1.19	0.87	1.12	0.69
Tobacco	0.95	1.05	1.08	1.18	1.28	1.17

Note that a single model is run within each site and therefore each hazard ratio is conditional on all other covariates in the model

Table 25 b. Adjusted hazard ratios for ischemic stroke by exposure of interest (RIVA) and confounders from site-specific Weibull accelerated failure time models

	Everyone		No Prior Ischemic Stroke		No Prior Cerebrovascular Disease	
	Site1	Site2	Site1	Site2	Site1	Site2
EXPOSURE						
RIVA	0.68	0.89	0.78	0.97	0.89	0.94
CONFOUNDERS						
Age (Ref:21-55)						
56-65	0.90	2.24	0.54	2.58	0.54	2.50
66-75	0.84	1.57	0.66	1.63	0.83	1.29
76+	1.11	2.44	1.06	2.48	1.13	2.03
Sex (Ref: Male)	1.26	1.20	1.39	1.17	1.37	1.37
Comorbidity Score (Ref:-2-0)						
1-4	1.05	1.25	1.08	1.06	1.13	0.95
≥5	1.54	2.31	1.78	1.80	1.96	1.73
1+ ER Visits	1.00	1.04	0.90	1.22	0.84	1.12
1+ Inpatient Visits	2.06	1.76	1.52	0.94	1.37	0.90
Year (Ref: 2013)						
2014	1.05	1.10	1.09	1.08	0.98	1.23
2015	1.06		1.09		0.85	
Cerebrovascular Disease	4.69	4.96	1.97	1.09		
Heart Failure/Cardiomyopathy	1.07	1.26	1.02	1.04	1.18	1.03
Hypertension	0.82	1.43	0.69	1.64	0.63	1.90
Hyperlipidemia	1.21	0.82	1.14	0.71	1.07	0.61
Coronary Artery Disease	0.78	1.09	0.87	1.62	1.02	1.70
Peripheral vascular disease	0.99	0.58	1.08	1.23	0.95	1.73
Diabetes	1.10	1.10	1.22	1.12	1.21	1.04
Renal Disease	1.01	0.68	1.19	0.87	1.12	0.69
Tobacco	0.95	1.06	1.08	1.18	1.28	1.17

Note that a single model is run within each site and therefore each hazard ratio is conditional on all other covariates in the model

Table 26. Adjusted hazard ratios for time to censoring by exposure of interest (RIVA) and confounders from site-specific Weibull accelerated failure time models

	No Prior Cerebrovascular Disease	
	Site 1	**Site2**
EXPOSURE		
RIVA	1.05	0.95
CONFOUNDERS		
Age (Ref:21-55)		
56-65	1.03	0.86
66-75	0.92	0.69
76+	0.92	0.69
Sex (Ref: Male)	1.02	0.99
Comorbidity Score (Ref:-2-0)		
4-Jan	1.03	0.97
≥5	1.14	1.06
1+ ER Visits	1.05	1.03
1+ Inpatient Visits	1.13	1.06
Year (Ref: 2013)		
2014	1.58	4.10
2015	8.87	
Heart Failure/Cardiomyopathy	0.99	1.00
Hypertension	0.97	0.95
Hyperlipidemia	0.99	0.90
Coronary Artery Disease	1.01	0.99
Peripheral vascular disease	1.02	0.96
Diabetes	1.03	1.00
Renal Disease	0.98	1.01
Tobacco	1.06	1.01

2. Simulation Generation and Evaluation Study

To mimic this Rivaroxaban and Ischemic Stroke real world data example, we generated simulated data using the framework detailed in **Section II**, but we will briefly summarize here. For each site we calculated the following summary statistics:

- Confounders: Probabilities within Confounders (**Table 23**) and Common Probabilities between each confounder category and all other confounders
- Exposure|Confounders (Propensity Model): Coefficients from a logistic model fitting the outcome RIVA versus WARF with all covariates in the model (**Table 24**)
- Outcome|Exposure and Confounders: Coefficients from a Weibull Accelerated Failure time for the outcome time to ischemic stroke including exposure and confounders in the model. We use the true coefficents for the confounder variables at each site (**Table 25 b**) and alter the exposure coefficients depending on the strength of relationship desired.
- Censored|Exposure, Confounders: Allowed for the three most prevalent modes in prescribing patterns (typically 30, 90, and 180 days, but we allowed the data to choose the most common modes, so they varied by site) and returned the prevalence of the modes and time of each mode (See **Figure 8** for common censoring modes by site). To model the censoring distribution amongst those that were not censored at any of the three most common prescribing modes we obtained coefficients from the Weibull Accelerated Failure time model for the outcome time to censoring (censor now the outcome) and we censored in this model at the time of ischemic stroke (ischemic stroke now the censor variable) (**Table 26**). This model was fit amongst only those that were not censored at the three most common prescribing modes and did not include covariates.

These summary statistics were then used to simulate datasets independently for each site, including simulated covariates, exposure and time-to-event or censoring. See **Section II** for details of how to simulate such data.

We performed 2,000 simulations for each of the four different treatment effect scenarios using total sample sizes of 40,000 distributed by the proportionate size of each site in the example datasets. The HR comparing Rivaroxaban with Warfarin was set to 1.0, 0.80 and 0.67 and was the same at each site (homogeneous). For each set of simulated data, all proposed models were fit and the resulting estimates, standard errors, test statistics and hypothesis tests returned. Estimates of bias (on the log HR scale), power (based on log rank test) and coverage (using Wald 95% CI on log HR scale) are presented in **Table 26**.

3. RIVA and Ischemic Stroke Simulation Study Results

Simulation results are presented in **Table 27**. Overall the results were favorable for all estimators. The pooled analysis methods – adjusting directly for confounders, adjusting for deciles of the propensity score or adjusting for the propensity score using B-splines – and stratification methods – stratifying on site and adjusting for confounders, or stratifying on deciles of the propensity score – were found to have comparable performance in terms of bias, type I error, power, and coverage. Type-I error was consistently lower than 0.05 for the pooled analysis relative to the site-specific propensity score analyses. However, contrary to the ACEI and angioedema example, we did not find a clear better performance in terms of bias when using 10 or 15 quantiles relative to 5 quantiles for pooled analyses. We did observe this expected trend when using site-specific propensity score stratification or adjustment indicating a clear need for more than 5 quantiles. Further, even in the pooled analysis the

performance was not noticeably worse when adjusting for additional quantiles. Therefore, we still recommend adjusting for more than 5 quantiles to assure confounding control without any large loss in power or coverage. Also, site-specific propensity score methods tended to perform better then pooled propensity score adjustment approaches.

Table 27. Simulation results with homogeneous effects across sites (2 sites, 2,000 simulations, samples of size 40,000)

	HR = 1.0			HR = 0.80			HR = 0.67		
Model	**Bias**	**Type I**	**Coverage**	**Bias**	**Power**	**Coverage**	**Bias**	**Power**	**Coverage**
Pooled Data									
Adj Confounders + Site	0.0008	0.042	0.955	0.0001	0.552	0.949	-0.0028	0.937	0.949
Adj PS Indicators									
5 quantiles	-0.0008	0.044	0.956	-0.0014	0.564	0.945	-0.0044	0.941	0.950
10 quantiles	0.0022	0.040	0.955	0.0017	0.550	0.948	-0.0012	0.937	0.948
15 quantiles	0.0029	0.042	0.955	0.0024	0.545	0.948	-0.0006	0.935	0.950
20 quantiles	0.0032	0.041	0.956	0.0026	0.543	0.948	-0.0003	0.935	0.950
Adj PS B-splines	0.0049	0.038	0.956	0.0044	0.537	0.951	0.0015	0.932	0.949
Stratify Site Adj Conf	0.0007	0.043	0.955	0.0001	0.555	0.948	-0.0027	0.937	0.948
Stratify PS									
5 quantiles	-0.0009	0.049	0.955	-0.0015	0.576	0.946	-0.0045	0.943	0.950
10 quantiles	0.0022	0.046	0.955	0.0017	0.561	0.947	-0.0012	0.939	0.948
15 quantiles	0.0029	0.045	0.955	0.0024	0.556	0.948	-0.0005	0.940	0.951
20 quantiles	0.0031	0.043	0.956	0.0027	0.557	0.950	-0.0003	0.938	0.950
Site-Specific									
Adj Site-PS Indicators									
5 quantiles	-0.0096	0.055	0.954	-0.0097	0.591	0.947	-0.0125	0.949	0.952
10 quantiles	-0.0033	0.047	0.953	-0.0035	0.570	0.948	-0.0065	0.940	0.950
15 quantiles	-0.0030	0.045	0.954	-0.0033	0.567	0.949	-0.0063	0.941	0.950
20 quantiles	-0.0027	0.046	0.953	-0.0030	0.562	0.949	-0.0062	0.940	0.950
Adj Site-PS B-splines	-0.0023	0.044	0.956	-0.0026	0.564	0.949	-0.0055	0.938	0.950
Stratify Site+Site-PS									
5 quantiles	-0.0096	0.058	0.955	-0.0096	0.602	0.947	-0.0124	0.950	0.952
10 quantiles	-0.0034	0.050	0.952	-0.0034	0.578	0.947	-0.0065	0.944	0.950
15 quantiles	-0.0031	0.048	0.954	-0.0032	0.575	0.949	-0.0061	0.944	0.951
20 quantiles	-0.0028	0.048	0.953	-0.0028	0.576	0.949	-0.0059	0.946	0.951
Stratify Site Adj B-splines	-0.0023	0.044	0.956	-0.0026	0.565	0.949	-0.0055	0.937	0.950
Reference Estimators Not for Methods Comparison									
Marginal Simulated	-0.0052			0.0021			-0.0033		
Unadjusted	-0.1095	0.000	0.849	-0.1088	0.000	0.867	-0.1120	0.000	0.863

V. STATISTICAL METHODS EXTENSIONS TO THE DISTRIBUTED DATA SETTING

There are several approaches to extend Cox PH regression and stratification to the distributed data setting. We will first discuss a method for de-identifying subject level data that aggregates event and censoring time into categories. For Cox PH stratification which does not adjust for confounders in the model you can actually estimate standard stratified Cox PH estimates using risk set information and therefore there is no difference between non-distributed and distributed application. The final approach we will discuss will apply a Mantel-Haenszel(2) type test statistics using site-specific regression models that may be more appropriate when site heterogeneity is expected.

A. COX PH METHODS WITH AGGREGATED TIME AND CONFOUNDERS OR PROPENSITY SCORES

The previous methods outlined in **Section III** which use Cox PH regression required subject data to conduct the analyses since we used continuous time to event or censoring as the outcome. We propose a simple approach to deidentify data by categorizing time to event or censoring instead of using continuous time information. This approach is viable for Sentinel both because of the rare event setting and because censoring time is actually naturally categorized based on prescribing patterns (e.g. 30 days or 90 days' supply of the prescribed medication). Therefore, categorizing censoring time into 30 day or shorter intervals likely retains the majority of the information available in the actual datasets. Further, when using Cox PH regression methods, the actual time of event is not needed, but the ordering of the times is the key quantity of interest (e.g. it does not matter that in our dataset we observe the second event in the dataset at day 8; what matters is that the event was the second event and the population at risk at that second event is known). The size of bins (e.g. 7 days or 30 days) should be large enough to contain at least one event, but small enough not to include too many events. Therefore, if we categorize time of event into small enough categories to limit the number of event ties, we are maintaining the key features for data analysis. Even if we induce a certain number of tied event times by categorizing event times, the influence of the tie is minimal because the risk set information is staying relatively stable (the majority of censoring occurs at longer intervals and censoring is the main influence on change of risk set information). More specifically, the estimated HR will change minimally when not recognizing that the events were not actually ties, because the only difference in the analysis with continuous data is that for the event that happened second, the previous tied event would have been removed from their risk set. Since risk sets are large, the induced tie results in a risk set changing from 49,999 to 50,000 patients, and therefore there is minimal influence in the actual estimate. However, we will account for ties to get the correct variance estimates using Efron's approach and therefore confidence intervals may be slightly larger using categorized time instead of continuous time. We will evaluate this issue in our simulation study to see if there is any issue with bias, type I error, or power when categorizing the information.

We will illustrate what is meant by aggregation of time and confounders for a specific dataset so that one can understand how we would implement this method in Sentinel. We will first define the individual level dataset that will be aggregated at each site to be shared across sites. First, divide the assumed two-year study period time into quarters of a year, and categorize each participant's start day into the quarter in which that person first enters the study. Often, we are interested in adjusting for time since prescription patterns/confounding may be different over time. Another reason to set up the data this way is to allow for the conduct of a surveillance study in which the analysis happens multiple times over a study period. Specifically, assume that study started on January 1, 2012. Then, any study participants who initially entered the study from January 1, 2012 through March 31, 2012 (e.g. date participant started taking the exposure or comparator medical product and met enrollment criteria) is assigned to

study quarter 1. Participants who entered the study from April 1, 2012 through June 30, 2012 are assigned to study quarter 2, and so on, up through study quarter 8.

Each participant has exposure status, *X*, and covariates such as site of enrollment (Site = 1, 2, or 3) and Age Category (Age (years) = 35-39, 40-44, 45-49, 50-54, 55-59, 60-65) at study entry. At the specified analysis time *a,* they have the outcome indicator $\delta_{si}^{c}(a) = I(E_{si} < C_{si} \cap E_{si} < T_{si}^{c}(a))$ which indicates if the participant experienced an outcome before they were censored or the current analysis time ended. At analysis time *a,* they also have the time to event or censoring variable ($T_{si}^{c}(a)$), defined as the minimum of the time to event, censoring, or analyses time, categorized into weekly categories. We will now walk through a test example of 10 participants at site 1 with 4 on comparator and 6 on exposure of interest, and will demonstrate how the dataset is created at analysis time June 30, 2012.

Table 28. Example subject-level dataset at a site

Enrollment Date	Site	Age	Exposure	Date of Outcome	Date of Censoring	Outcome $\delta_{si}(a)$	Outcome Time $T_{si}(a)$
Jan 10, 2012	1	47	0	.	.	0	172
Feb 1, 2012	1	55	1	.	Mar 20, 2012	0	48
Feb 20, 2012	1	60	0	Apr 10, 2012	.	1	50
Mar 12, 2012	1	64	0	.	.	0	110
Mar 31, 2012	1	58	1	.	Apr 18, 2012	0	18
Apr 25, 2012	1	46	1	May 1, 2012	.	1	6
May 30, 2012	1	42	1	Jun 12, 2012	.	1	13
Jun 3, 2012	1	64	0	.	.	0	27
Jun 10, 2012	1	38	1	.	.	0	20
June 29, 2012	1	39	1	.	.	0	1

The first step is to deidentify the subject-level data in **Table 28** by creating categories for study quarter and age, and to calculate weeks from study start for Outcome Time, $T_{si}^{c}(a)$ as follows:

Table 29. Example subject-level deidentified dataset at a site

Study Qtr	Site	Age Cat	Exposure	Outcome $\delta_{si}^{c}(a)$	Outcome Time $T_{si}^{c}(a)$ in weeks
1	1	3	0	0	25
1	1	5	1	0	7
1	1	6	0	1	8
1	1	6	0	0	16
1	1	5	1	0	3
2	1	3	1	1	1
2	1	2	1	1	2
2	1	6	0	0	4
2	1	1	1	0	3
2	1	1	1	0	1

The next step is to aggregate the subject-level data so that several participants can be represented in each row, to provide deidentification and the smallest number of data rows possible. To do this we propose the following aggregate dataset:

Table 30. Example deidentified aggregate dataset at site

Study Qtr	Site	Age Cat	*N*	N_x	*Y*	Y_x	E_1^0	E_2^0	…	E_8^0	…	E_{25}^0	C_1^0	C_2^0	C_3^0	C_4^0	…	C_{16}^0	…	C_{25}^0
1	1	3	1	0	0	0	0	0		0		0	0	0	0	0		0		1
1	1	5	2	2	0	0	0	0		0		0	0	0	0	0		0		0
1	1	6	2	0	1	0	0	0		1		0	0	0	0	0		1		0
2	1	1	2	2	0	0	0	0		0		0	0	0	0	0		0		0
2	1	2	1	1	1	1	0	0		0		0	0	0	0	0		0		0
2	1	3	1	1	1	1	0	0		0		0	0	0	0	0		0		0
2	1	6	1	0	0	0	0	0		0		0	0	0	0	1		0		0

E_1^1	E_2^1	E_3^1	E_4^1	…	E_{25}^1	C_1^1	C_2^1	C_3^1	…	C_7^1	…	C_{25}^1
0	0	0	0		0	0	0	0		0		0
0	0	0	0		0	0	0	1		1		0
0	0	0	0		0	0	0	0		0		0
0	0	0	0		0	1	0	1		0		0
0	1	0	0		0	0	0	0		0		0
1	0	0	0		0	0	0	0		0		0
0	0	0	0		0	0	0	0		0		0

where within each row defining study quarter and confounder stratum, we define the following counts: *N* is total number, N_x is the number exposed, *Y* is the total number of outcomes, Y_x is the number of exposed outcomes, E_W^0 is the number of outcomes in the comparator group observed at $T_{si}^c(a)$=*w*, C_W^0 is the number censored in the comparator group observed at $T_{si}^c(a)$=*w*, E_W^1 is the number of outcomes in the exposed group observed at $T_{si}^c(a)$=*w*, and C_W^1 is the number censored in the exposed group observed at $T_{si}^c(a)$=*w*. The number of rows in the dataset will be at most the number of study quarters times the number of confounder categories. As the sample size increases, the number of rows in the dataset will not increase beyond this maximum. This dataset can be securely sent to the coordinating center where the data can be de-aggregated to form the dataset needed to conduct the analysis.

This aggregation method can also be used for propensity score indicators instead of specific confounder strata. Therefore, it can be implemented both for Cox PH with adjustment for confounders and site directly (Adj Confounders+Site) and Cox PH with adjustment for site specific propensity score indicators (Adj Site-PS Indicators).

B. SITE AND SITE-SPECIFIC PROPENSITY SCORE-STRATIFIED COX PH REGRESSION

Conducting stratified Cox PH regression in which the strata are defined as site and site-specific propensity score percentile strata (Stratify Site+Site-PS) does not actually require subject level data to be shared across sites. The only information required to be shared across sites is a single row per event with the following deidentified information: site-specific propensity score percentile strata, if the event was exposed or the comparator group, ordering of event time within each stratum (1st event, 2nd event, and so on but not the actual event time) and the risk set at that time of the analysis. Specifically, the dataset needed would be the following:

Table 31. Example of a stratified regression dataset

Site	PS Stratum	Exposure	Event Order w/in Stratum	Number in Risk Set
1	1	0	1	20000
1	1	1	2	19500
1	1	0	3	16000
1	2	0	1	18000
1	2	1	2	17999
1	3	1	1	20000
1	3	1	2	18000
1	3	0	3	12000
1	3	1	4	2000
1	4	1	5	19500

This is the same information used when conducting the analysis using the continuous event and censoring time information so this is not a new method but just an approach to simplify the data returned.

C. MANTEL-HAENSZEL TYPE TEST STATISTIC IN DISTRIBUTED DATA SETTING

To limit data transmission, an alternative to categorizing all confounders and time is for each site to run a site-specific model and to transmit centrally only summary statistics. Specifically, we will estimate site-specific Cox PH models and use the HR site-specific estimate, $\hat{\Delta}_s$, and calculate an overall estimate, $\hat{\Delta}$, which is

$$\hat{\Delta} = \frac{\sum_{s=1}^{S} w_s \hat{\Delta}_s}{\sum_{s=1}^{S} w_s},$$

with estimated variance

$$\hat{\mathrm{V}}(\hat{\Delta}) = \frac{\sum_{s=1}^{S} w_s^2 \hat{\mathrm{V}}(\hat{\Delta}_s)}{\left[\sum_{s=1}^{S} w_s\right]^2},$$

where w_s can be the sample size of the site, N_s, or the inverse of the variance of the estimator from that site, $\hat{V}(\hat{\Delta}_s)$. For the simulation study we assess the following two MH models in which we fit site Cox PH models adjusting for covariates directly or using the cubic B-Splines:

MH Inv Var: $\lambda(T_{si},\delta_{si}|X_{si},\boldsymbol{Z}_{si}) = \lambda_0(T_{si})exp\left[\beta_{S,X}^{Adj}X_{si} + \boldsymbol{\beta}_z\,\boldsymbol{Z}_{si}\right]$ with $w_s = 1/\hat{V}(\hat{\beta}_{S,X}^{Adj})$ (11)

MH B-Splines Inv Var: $\lambda(T_{si},\delta_{si}|X_{si},\boldsymbol{e}^{*}_{si}) = \lambda_0(T_{si})exp\left[\beta_{S,X}^{BS}X_{si} + \boldsymbol{\beta}_p^s f(\boldsymbol{e}^{*}_{si})\right]$ with $w_s = 1/\hat{V}(\hat{\beta}_{S,X}^{BS})$ (12)

The only information necessary to send across sites is sample size, adjusted HR, and variance of the adjusted HR. Further, it would be preferable to also submit table 1 type information which includes the sample size, number of outcomes and total follow-up time by exposure and confounder categories.

Table 32. Summary of distributed methods evaluated

Method	Confounder Control	Confounder Sharing	De-identify
Adj Confounders+Site(1)	Regression on Categorical Confounders and Site	Pooled	Aggregate Time and Confounders
Adj Site-PS Indicators (5)	Regression on Site-Specific Propensity Scores (includes confounders only) Indicators and adjust for site and interactions with site	Site-Specific	Aggregate Time and PS-Indicators
Stratify Site+Site-PS (10)	Stratify on Site and Site-Specific Propensity Scores (includes confounders only) categories	Site-Specific	Risk Sets
MH Inv Var (11)	At site regress on Categorical Confounders and estimate overall HR using Mantel-Haenzel approach weighting on the inverse variance of the log HR	Site-Specific	Summary Info
MH B-Splines Inv Var(12)	At site regress on Site-Specific Propensity Score (includes confounders only) B-splines and estimate overall HR using Mantel-Haenzel approach weighting on the inverse variance of the log HR	Site-Specific	Summary Info

* Numbers refer to equation number referenced earlier in the report

VI. SIMULATION EVALUATION FOR THE DISTRIBUTED DATA SETTING

For this simulation evaluation we will use all of the scenarios described previously for Non-Distributed data in **Sections IV B** and **C**.

A. SIMULATION DISTRIBUTION OF PROPENSITY SCORE COEFFICIENTS FROM SITE 5

We will use the ACEI and Angioedema simulation scenario previously described in detail in **Section IV B**. We performed 2,000 simulations for each of the four different treatment effect scenarios using total sample sizes of 150,000 distributed by the proportionate size of each site in the example datasets. In the first three scenarios, the HR comparing ACEI with BB was set to 1.0, 1.5 and 2.0 and was the same at each site (homogeneous). The fourth scenario allowed the HR to be heterogeneous/vary by site in the same way that the estimates varied in the observed example data. To calculate the pooled HR estimate in the setting where the HR is heterogeneous we fit a site stratified Cox PH model with all confounders and a single term for the effect of exposure. This estimate of the HR was used as the truth for the simulations that included heterogeneity of the effect of exposure on outcome. For each set of simulated data, all proposed distributed data methods were fit and the resulting estimates, standard errors, test statistics and hypothesis tests returned. Estimates of bias (on the log HR scale), power and coverage are presented in **Table 33** and **Table 34**. We further re-conducted the same simulation study with the exception of removing the very small site 5 (2.8% of the total study population) with results presented in **Table 35** and **Table 36.**

Simulation results are presented in **Table 33** through **Table 36**. Overall the results were favorable for all estimators, with the exception of the Meta Analytic MH type methods. As shown in **Table 33**, the bias was consistently the largest for the MH estimators especially when there was an elevated treatment effect (HR=1.5 or 2.0), both when the site-specific models directly adjusted for covariates and when they adjusted for the propensity score using B-splines. Note that we included the use of B-splines in an attempt to reduce the number of parameters being estimated in site-specific models, and thereby reduce the bias of the MH estimators. The MH estimator's type I error was consistently low, and coverage was above 95%. For example, in **Table 33** the MH method when adjusting for covariates had a type I error of 2.9% which is statistically different from 5% (p=0.03). To assess whether the estimator's poor performance was in part due to the relative imbalance caused by having one very small data partner contributing to the analysis (Site 5 had only 2.8% of the total study population), we conducted identical simulations excluding Site 5 (**Table 35** and **Table 36**). Removing the very small site improved the performance of the MH estimators, but compared to the other methods evaluated here, MH estimators still performed relatively poorly.

When the treatment effect is homogeneous among the sites (**Table 33** and **Table 35**), pooled analysis methods and distributed methods performed very comparable in terms of bias, type I error, power, and coverage. They performed so well that estimates were almost the same as having continuous time to event information relative to de-identifying censoring and outcome time into 7 or 30-day intervals. This is likely due to the rare event setting as well as censoring mainly happening on fixed interval times (e.g. 30-day prescription fills) so information loss is minimalized with additional aggregation of the information. Therefore, given this scenario is common in Sentinel we would recommend using this de-identified aggregation approach since it allows for information sharing and analysis flexibly like subgroups to be conducted without loss of information.

Another key finding was that using quintiles of the propensity score may provide insufficient control of confounding, whether used for adjustment or stratification. Therefore, caution should be taken in the

Sentinel context when categorizing propensity score. We did not observe less bias when we further partitioned data into 15 or 20 quantiles for the homogeneous treatment effect case. Nominal coverage was achieved by all estimators with the exception of the MH-type estimators as previously noted.

In the setting with small amounts of site heterogeneity (**Table 34** and **Table 36**) in which site-specific models should theoretically outperform pooled data models, we did not find an appreciable difference. This likely reflects the moderate differences observed across sites: HRs of 2.40, 3.62, 3.39, 2.96, and 2.39, respectively and is a limitation of the example we are using for the interim report. However, we did notice some minor improvements when using 15 quantiles relative to 10 quantiles in this setting (**Table 34)**, but this improvement was not observed after removing the small site (**Table 36**). Therefore, 10 quantiles may be sufficient, but sensitivity analyses looking at 15 or 20 quantiles may also be recommended.

Table 33. Simulation results with homogeneous effects across sites (5 sites, 2,000 simulations, samples of size 150,000)

	HR = 1.0			HR = 1.5			HR = 2.0		
	Bias	Type I	Coverage	Bias	Power	Coverage	Bias	Power	Coverage
Pooled Data									
Adj Confounders + Site									
Continuous	0.002	0.039	0.961	0.002	0.673	0.961	0.006	0.989	0.959
7-Day Time Interval	0.002	0.039	0.961	0.002	0.673	0.961	0.006	0.989	0.959
30-Day Time Interval	0.002	0.039	0.961	0.002	0.673	0.961	0.006	0.989	0.959
Site Specific									
Adj Site-PS Indicators									
5 Quantiles									
Continuous	-0.005	0.037	0.957	-0.006	0.662	0.959	-0.003	0.988	0.956
7-Day Time Interval	-0.005	0.037	0.957	-0.006	0.661	0.959	-0.003	0.988	0.956
30-Day Time Interval	-0.005	0.037	0.957	-0.006	0.661	0.959	-0.003	0.988	0.956
10 Quantiles									
Continuous	-0.001	0.040	0.959	0.000	0.664	0.961	0.003	0.988	0.957
7-Day Time Interval	-0.001	0.040	0.959	0.000	0.664	0.961	0.003	0.988	0.957
30-Day Time Interval	-0.001	0.040	0.959	0.000	0.665	0.960	0.003	0.988	0.957
15 Quantiles									
Continuous	0.003	0.041	0.961	0.004	0.670	0.960	0.007	0.989	0.958
7-Day Time Interval	0.003	0.041	0.961	0.004	0.670	0.960	0.007	0.989	0.958
30-Day Time Interval	0.003	0.041	0.961	0.004	0.670	0.959	0.007	0.989	0.958
20 Quantiles									
Continuous	0.003	0.040	0.959	0.004	0.663	0.959	0.007	0.979	0.956
7-Day Time Interval	0.003	0.040	0.959	0.004	0.663	0.959	0.007	0.979	0.956
30-Day Time Interval	0.003	0.040	0.960	0.004	0.662	0.959	0.007	0.979	0.956
Stratify Site + Site-PS									
5 Quantiles	-0.005	0.040	0.957	-0.006	0.681	0.958	-0.003	0.989	0.956
10 Quantiles	-0.001	0.046	0.959	0.000	0.680	0.958	0.003	0.990	0.957
15 Quantiles	0.003	0.046	0.961	0.003	0.688	0.959	0.007	0.989	0.956
20 Quantiles	0.003	0.046	0.960	0.004	0.689	0.958	0.007	0.990	0.955
MH Inv. Variance	-0.004	0.029	0.966	-0.014	0.610	0.966	-0.018	0.976	0.963
MH BS Inv. Variance	-0.004	0.030	0.966	-0.014	0.610	0.967	-0.019	0.975	0.960
Reference Estimators Not for Methods Comparison									
Marginal Simulated	0.007		0.958	0.001		0.954	-0.002		0.945
Unadjusted	-0.094		0.933	-0.094		0.923	-0.090		0.915

Table 34. Simulation results with observed treatment heterogeneity across sites (5 sites, 2000 simulations)

	HR = 2.94		
	Bias	**Power**	**Coverage**
Pooled Data			
Adj Confounders + Site			
Continuous	-0.007	1.000	0.952
7-Day Time Interval	-0.007	1.000	0.952
30-Day Time Interval	-0.007	1.000	0.952
Site Specific			
Adj Site-PS Indicators			
5 Quantiles			
Continuous	-0.016	1.000	0.948
7-Day Time Interval	-0.016	1.000	0.948
30-Day Time Interval	-0.016	1.000	0.948
10 Quantiles			
Continuous	-0.010	1.000	0.949
7-Day Time Interval	-0.010	1.000	0.949
30-Day Time Interval	-0.010	1.000	0.949
15 Quantiles			
Continuous	-0.006	1.000	0.951
7-Day Time Interval	-0.006	1.000	0.950
30-Day Time Interval	-0.006	1.000	0.950
20 Quantiles			
Continuous	-0.005	0.993	0.952
7-Day Time Interval	-0.005	0.993	0.952
30-Day Time Interval	-0.005	0.993	0.952
Stratify Site + Site-PS			
5 Quantiles	-0.016	1.000	0.947
10 Quantiles	-0.010	1.000	0.949
15 Quantiles	-0.006	1.000	0.951
20 Quantiles	-0.006	1.000	0.952
MH Inv. Variance	-0.041	1.000	0.944
MH BS Inv. Variance	-0.041	1.000	0.943
Reference Estimators Not for Methods Comparison			
Marginal Stratified	0.000		
Unadjusted	-0.095		0.895

Table 35. Simulation results with homogeneous effects across sites, after excluding the smallest site (4 sites, 2,000 simulations, samples of size 150,000)

	HR = 1.0			HR = 1.5			HR = 2.0		
	Bias	Type I	Coverage	Bias	Power	Coverage	Bias	Power	Coverage
Pooled Data									
Adj Confounders + Site									
Continuous	0.002	0.045	0.957	-0.007	0.635	0.948	0.001	0.989	0.953
7-Day Time Interval	0.002	0.045	0.958	-0.007	0.635	0.948	0.001	0.989	0.953
30-Day Time Interval	0.002	0.045	0.957	-0.007	0.635	0.948	0.001	0.989	0.954
Site Specific									
Adj Site-PS Indicators									
5 Quantiles									
Continuous	-0.005	0.041	0.961	-0.015	0.621	0.946	-0.006	0.991	0.951
7-Day Time Interval	-0.005	0.041	0.961	-0.015	0.621	0.946	-0.006	0.991	0.951
30-Day Time Interval	-0.005	0.041	0.960	-0.015	0.621	0.946	-0.006	0.991	0.951
10 Quantiles									
Continuous	0.000	0.043	0.961	-0.010	0.635	0.946	-0.001	0.992	0.954
7-Day Time Interval	0.000	0.043	0.961	-0.010	0.634	0.946	-0.001	0.992	0.954
30-Day Time Interval	0.000	0.044	0.961	-0.010	0.634	0.946	-0.001	0.992	0.954
15 Quantiles									
Continuous	0.004	0.047	0.960	-0.006	0.638	0.948	0.003	0.991	0.953
7-Day Time Interval	0.004	0.047	0.960	-0.006	0.638	0.948	0.003	0.991	0.953
30-Day Time Interval	0.004	0.047	0.960	-0.006	0.638	0.948	0.003	0.991	0.953
20 Quantiles									
Continuous	0.004	0.043	0.959	-0.006	0.637	0.949	0.003	0.981	0.954
7-Day Time Interval	0.004	0.043	0.959	-0.006	0.638	0.949	0.003	0.981	0.954
30-Day Time Interval	0.004	0.043	0.959	-0.006	0.636	0.949	0.003	0.981	0.954
Stratify Site + Site-PS									
5 Quantiles	-0.005	0.045	0.961	-0.015	0.640	0.945	-0.007	0.992	0.951
10 Quantiles	0.000	0.049	0.961	-0.010	0.652	0.947	-0.001	0.992	0.954
15 Quantiles	0.004	0.053	0.962	-0.006	0.658	0.948	0.003	0.993	0.955
20 Quantiles	0.004	0.054	0.960	-0.006	0.659	0.949	0.003	0.992	0.955
MH Inv. Variance	-0.001	0.036	0.965	-0.020	0.588	0.956	-0.016	0.982	0.958
MH BS Inv. Variance	-0.001	0.034	0.965	-0.020	0.586	0.956	-0.016	0.982	0.958
Reference Estimators Not for Methods Comparison									
Marginal Simulated	0.000		0.955	0.007		0.952	0.010		0.955
Unadjusted	-0.090		0.929	-0.100		0.914	-0.093		0.914

Table 36. Simulation results with observed treatment heterogeneity across sites after excluding the smallest site (4 sites, 2000 simulations, samples of size 150,000)

	HR = 2.94		
	Bias	**Power**	**Coverage**
Pooled Data			
Adj Confounders + Site			
Continuous	0.001	1.000	0.945
7-Day Time Interval	0.001	1.000	0.945
30-Day Time Interval	0.001	1.000	0.945
Site Specific			
Adj Site-PS Indicators			
5 Quantiles			
Continuous	-0.007	1.000	0.945
7-Day Time Interval	-0.007	1.000	0.945
30-Day Time Interval	-0.007	1.000	0.947
10 Quantiles			
Continuous	-0.002	1.000	0.945
7-Day Time Interval	-0.002	1.000	0.946
30-Day Time Interval	-0.002	1.000	0.946
15 Quantiles			
Continuous	0.002	1.000	0.945
7-Day Time Interval	0.002	1.000	0.945
30-Day Time Interval	0.002	1.000	0.945
20 Quantiles			
Continuous	0.002	0.991	0.947
7-Day Time Interval	0.002	0.991	0.947
30-Day Time Interval	0.002	0.991	0.946
Stratify Site + Site-PS			
5 Quantiles	-0.007	1.000	0.946
10 Quantiles	-0.002	1.000	0.946
15 Quantiles	0.002	1.000	0.946
20 Quantiles	0.002	1.000	0.947
MH Inv. Variance	-0.035	1.000	0.943
MH BS Inv. Variance	-0.035	1.000	0.944
Reference Estimators Not for Methods Comparison			
Marginal Stratified	0.000		
Unadjusted	-0.090		0.907

B. SIMULATION STUDY FOR RIVAROXABAN AND ISCHEMIC STROKE EVALUATION

We will use the Rivaroxaban and Ischemic stroke simulation scenario previously described in detail in Section IV C. We performed 2,000 simulations for each of the four different treatment effect scenarios using total sample sizes of 40,000 1distributed by the proportionate size of each site in the example datasets. The HR comparing RIVA with WARF was set to 1.0, 0.80 and 0.67 and was the same at each site (homogeneous). For each set of simulated data, all proposed distributed data methods were fit and the resulting estimates, standard errors, test statistics and hypothesis tests returned. Estimates of bias (on the log HR scale), power and coverage are presented in **Table 37.**

Overall the results were favorable for all estimators, but the Meta Analytic MH type methods had slightly less power than the other approaches. However, bias was not an issue for the MH type methods as was shown in the other Angioedema and ACEI example. This is likely due to only having two sites that were also not as different in sample size relative to the other example.

We found that again pooled analysis methods and distributed methods performed comparably in terms of bias, type I error, power, and coverage. They performed so well that estimates were almost the same as having continuous time to event information relative to de-identifying censoring and outcome time into 7 or 30-day intervals. This is likely due to the rare event setting as well as censoring mainly happening at fixed interval times (e.g. 30-day prescription fills) so information loss is minimized with additional aggregation of the information. Therefore, given that this scenario is common in Sentinel, we would recommend using this de-identified aggregation approach since it allows for information sharing. Another advantage is that subgroup analyses can be easily conducted given one includes the subgroup covariate in the returned dataset. Other deidentification approaches such as PS stratification which shares risk set information stratified by PS stratum and MH type methods do not provide the information to easily allow subgroup analyses to be conducted centrally on the same datasets.

Another key finding was that using quintiles of the propensity score may provide insufficient control of confounding, whether used for adjustment or stratification. Therefore, caution should be taken in the Sentinel context when categorizing propensity score. We observed less bias when we further partitioned data into 15 or 20 quantiles. This was a slightly different finding than in the ACEI and angioedema example and therefore provides some recommendation for minimally using 10 quantiles with further sensitivity analyses for 15 or 20 quantiles.

Table 37. Simulation results with homogeneous effects across sites (2 sites, 2,000 simulations, samples of size 40,000)

	HR = 1.0			HR = 0.80			HR = 0.67		
	Bias	Type I	Coverage	Bias	Power	Coverage	Bias	Power	Coverage
Pooled Data									
Adj Confounders + Site									
Continuous	0.0008	0.042	0.955	0.0001	0.552	0.949	-0.0028	0.937	0.949
7-Day Time Interval	0.0008	0.042	0.955	0.0001	0.553	0.949	-0.0028	0.937	0.950
30-Day Time Interval	0.0007	0.041	0.955	0.0002	0.550	0.949	-0.0027	0.936	0.951
Site Specific									
Adj Site-PS Indicators									
5 Quantiles									
Continuous	-0.0096	0.055	0.954	-0.0097	0.591	0.947	-0.0125	0.949	0.952
7-Day Time Interval	-0.0096	0.055	0.954	-0.0097	0.591	0.947	-0.0125	0.948	0.952
30-Day Time Interval	-0.0096	0.054	0.955	-0.0096	0.590	0.948	-0.0125	0.948	0.952
10 Quantiles									
Continuous	-0.0033	0.047	0.953	-0.0035	0.570	0.948	-0.0065	0.940	0.950
7-Day Time Interval	-0.0033	0.047	0.953	-0.0035	0.570	0.948	-0.0065	0.941	0.950
30-Day Time Interval	-0.0033	0.046	0.953	-0.0034	0.567	0.948	-0.0065	0.940	0.950
15 Quantiles									
Continuous	-0.0030	0.045	0.954	-0.0033	0.567	0.949	-0.0063	0.941	0.950
7-Day Time Interval	-0.0030	0.045	0.954	-0.0033	0.566	0.949	-0.0063	0.940	0.950
30-Day Time Interval	-0.0030	0.045	0.954	-0.0033	0.563	0.949	-0.0062	0.941	0.951
20 Quantiles									
Continuous	-0.0027	0.046	0.953	-0.0030	0.562	0.949	-0.0062	0.940	0.950
7-Day Time Interval	-0.0027	0.046	0.953	-0.0030	0.564	0.949	-0.0062	0.940	0.950
30-Day Time Interval	-0.0027	0.045	0.953	-0.0029	0.561	0.949	-0.0061	0.940	0.951
Stratify Site + Site-PS									
5 Quantiles	-0.0096	0.058	0.955	-0.0096	0.602	0.947	-0.0124	0.950	0.952
10 Quantiles	-0.0034	0.050	0.952	-0.0034	0.578	0.947	-0.0065	0.944	0.950
15 Quantiles	-0.0031	0.048	0.954	-0.0032	0.575	0.949	-0.0061	0.944	0.951
20 Quantiles	-0.0028	0.048	0.953	-0.0028	0.576	0.949	-0.0059	0.946	0.951
MH Inv. Variance	-0.0005	0.048	0.955	0.0000	0.537	0.948	-0.0018	0.930	0.950
MH BS Inv. Variance	-0.0003	0.045	0.956	0.0003	0.538	0.950	-0.0014	0.930	0.951
Reference Estimators Not for Methods Comparison									
Marginal Simulated	-0.0052			0.0021			-0.0033		
Unadjusted	-0.1095	0.000	0.849	-0.1088	0.000	0.867	-0.1120	0.000	0.863

VII. DISCUSSION AND CONCLUSIONS

In this final report we have proposed and evaluated 12 different Cox PH survival methods to adjust for confounding including direct confounder adjustment and numerous variants of propensity score adjustment and stratification. The methods that performed the best included direct adjustment, propensity score adjustment with 10 indicator strata or more, propensity score adjustment using b-splines, and propensity score stratification with 10 strata or more. We further showed that using site-specific propensity scores performed equally well or better than fitting an overall propensity score. Therefore, since site-specific propensity scores are both more feasible in Sentinel (distributed data) and more scientifically appropriate since sites likely have different prescription patterns yielding different exposure cohorts, having equivalent or better performance is promising.

We further found that extending methods to the distributed data setting by aggregating censoring and outcome time performed as well as non-distributed methods using continuous censoring and outcome time. This finding occurred for several reasons. First, we are applying Cox PH methods which are time invariant and only take into account time by order of outcome events and risk sets available at the time of the event. Therefore, if an outcome occurs at day 35 it only categorizes data as being available for the risk set after day 35. The method will give you the same result if someone was censored at day 36 as if they were censored at day 40 given no new outcomes occurred between day 36 and 40. Since we are in both in the rare event setting and censoring mainly happens at fixed time intervals (e.g. 30-day prescription fills) information loss is minimized with additional aggregation of the information and does not actually change the estimated HR that strongly and often not at all. Further, since risk sets are large, given most participants do not have an event, misclassifying a handful of observations as being in the risk set does not meaningfully change the denominator and therefore the resulting HR is not noticeably affected. In Sentinel, if you are applying Cox PH methods with rare events de-identifying data into 7-day time intervals is a simple and viable approach for conducting analyses.

There were several limitations to the simulation evaluation. We only mimicked two medical product comparisons which may be limited in generalizability. For the first example dataset, ACEI compared to BB on the outcome angioedema, ACEI, the exposure of interest, had actually been on the market for a significant amount of time when the data was pulled. Thus, ACEI use was more common than a new medical product would normally be. We chose this comparison since it was a known positive association between elevated rates of Angioedema and ACEI that was published, and data were readily available to conduct the simulation evaluation. We further included another example which was RIVA compared to WARF on the outcome ischemic stroke. RIVA is a new medical product which was an advantage. Ischemic stroke in this population was relatively common which tends to allow for all methods to perform more comparably. Another limitation of both examples for survival analysis in particular was that most participants took the medication for a short amount of time (30 to 90 days). Potentially for this example a binary outcome analysis may have been more appropriate especially since most of the effect occurs shortly after exposure. An advantage of this type of shorter term exposure was that it helped us think about the effects of censoring and how it should be mimicked in a simulation evaluation. That is why we added censoring bumps to allow for prescribing patterns that are likely to be observed in future Sentinel studies.

We also developed a new data simulation approach that mimics complex data using summary information. It performed equally as well to bootstrapping and other approaches that would require subject data to implement. This promising approach can be used by others in Sentinel and outside networks to simulate complex data with minimal data sharing.

Overall, this task order found that stratifying on site-specific propensity scores and adjusting for site-specific propensity scores are methods that perform well in terms of bias, type I error, power, and coverage. When applying these approaches, we recommend at least 10 quantiles of the propensity score and to conduct sensitivity analyses for 15 or 20 quantiles. We further presented methods tailored to the distributed data setting that performed as well as pooled analysis methods. Therefore, these propensity score methods are viable to the Sentinel distributed data network and will be straightforward to incorporate into the system.

VIII. REFERENCES

1. Mantel N, Haenszal W. Statistical aspects of the analysis of data from retrospective studies. J Natl Cancer Inst. 1959;22(4):719-48.
2. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22(4):719-48.
3. Cox DR. Regression Models and Life-Tables. JRSS, Series B. 1972;34:187-220.
4. Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A review. Sankha: The Indian Journal of Statistics, Series A. 1973:417-46.
5. de Boor C. A Practical Guide to Splines. New york: Springer-Verlag; 1978.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
7. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. JASA. 1984;79(387):516-24.
8. Rosenbaum PR, Rubin DB. Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. . The American Statistician. 1985;39(1):33-8.
9. Robins JM. A new approach to causal inference in mortality studies with a sustained expo- sure periodapplication to control of the healthy worker survivor effect. Mathematical Modeling. 1986;7(9):1393-512.
10. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chronic Diseases. 1987;40:139S-61S.
11. Slater EE, Merrill DD, Guess HA, Roylance PJ, Cooper WD, Inman WH, Ewan PW. Clinical profile of angioedema associated with angiotensin converting-enzyme inhibition. JAMA. 1988;260(7):967-70.
12. Potter FJ. A study of procedures to identify and trim extreme sampling weights. Proceedings of the section on survey research methods, American Statistical Association. 1990(225-230).
13. Frydenberg M. Marginalization and collapsibility in graphical interaction models. Annals of Statistics. 1990;18(2):790-805.
14. Hastie TJ. Generalized Additive Models. Chambers JM, Hastie TJ, editors: Wadsorth & Brooks/Cole; 1992.
15. Israili ZH, Hall WD. Cough and angioneurotic edema associated with angiotensin-converting enzyme inhibitor therapy. A review of the literature and pathophysiology. Ann Intern Med. 1992;117:234-42.
16. Potter FJ. The effect of weight trimming on nonlinear survey estimates. Proceedings of the American Statistical Association, Section on Survey Research Methods. 1993;758763.
17. Cheng P. Nonparametric estimation of mean functionals with data missing at random. JASA. 1994;89(425):81-7.
18. Tjøstheim D, Auestad BH. Nonparametric identification of nonlinear time series: projections.`. JASA. 1994;89(428):1398-409.
19. Newey WK. Kernel estimation of partial means and a general variance estimator. Econometric Theory. 1994;10(02):1-21.
20. Linton O, Nielsen JP. A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika. 1995;82(1):93-100.
21. Chambers RL, editor. Weighting and calibration in sample survey estimation. Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth 1997.

22. Sabroe RA, Black AK. Angiotensin-converting enzyme (ACE) inhibitors and angio-oedema. Br J Dermatol. 1997;136:153-8.
23. Leisch F, Weingessel A, Hornik K. On the generation of correlated artificial binary data. Working Paper Series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science": Vienna University of Economics; 1998 Contract No.: Document Number|.
24. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66(2):315-31.
25. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statistical Science. 1999;14(1):29-46.
26. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;550-560.
27. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962-73.
28. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med. 2005;24(11):1713-23.
29. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. J Clinical Epi. 2007;60(9):874-82.
30. Austin PC, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study. Stat Med. 2007;26(4):754-68.
31. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078-94.
32. Müller UU. Estimating linear functionals in nonlinear regression with responses missing at random. Annals of Statistics. 2009;37:2245-77.
33. Brown JS, Kulldorff M, Petronis KR, Reynolds R, Chan KA, Davis RL, Graham D, Andrade SE, Raebel MA, Herrinton L, Roblin D, Boudreau D, Smith D, Gurwitz JH, Gunter MJ, Platt R. Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. Pharmacoepidemiol Drug Saf. 2009;18(3):226-34.
34. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. AJE. 2010;172(9):1092-7.
35. Kaiser S, Träger D, Leisch F. Generating Correlated Ordinal Random Values: Department of Statistics, University of Munich; 2011. Report No.: Number 94 Contract No.: Document Number|.
36. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. . Pharmaceutical Statistics. 2011;10:150-61.
37. Wicklin R. Simulating Data with SAS. Cary, NC: SAS Institute Inc; 2013.
38. Hahn J, Ridder G. Asymptotic variance of semiparametric estimators with generated regressors. Econometrica. 2013;81(1):315-40.
39. Rothman KJ, Mosquin PL. Sparse-data bias accompanying overly fine stratification in an analysis of beryllium exposure and lung cancer risk. Annals of Epidemiology. 2013;23:43-8.
40. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med. 2013;32(16):2837-49. PMCID: PMC3747460.
41. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Stat Med. 2014;33:1242-58.

42. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014;72:219-26. PMCID: PMC3935334.
43. Carnahan R, Gagne J, Nelson J, Fireman B, Zhang R, Levenson M, Graham D, Tiwari RC. PROMPT: Rivaroxaban Surveillance Plan. Sentinel Ini; 2015 [updated 2015; cited]; Available from: https://www.sentinelinitiative.org/drugs/assessments/prompt-rivaroxaban-surveillance-plan.
44. Cook AJ, Wellman RD, Shoaibi A, Tiwari RC, Heckbert SR, Li L, Izem R, Zhang R, Nelson JC. Safety Signaling Methods for Survival Outcomes to Control for Confounding in the MSDD. 2015 [updated 2015; cited]; Available from: https://www.sentinelsystem.org/sites/default/files/Drugs/Assessments/Mini-Sentinel_Methods_Safety-Signaling-Methods_Survival-Outcomes_Confounding-MSDD.pdf.
45. Nelson JC, Boudreau D, Wellman R, Yu O, Cook AJ, Maro J, Ouellet-Hellstrom R, Floyd J, Heckbert SR, Pinheiro S, Reichman M, Shoaibi A. Improving Sequential Safety Surveillance Planning Methods for Routine Assements that use Regression Adjustment or Weighting to Control Confounding. 2016 [updated 2016; cited]; Available from: https://www.sentinelsystem.org/sentinel/methods/routine-prospective-safety-surveillance-new-drugs-vaccines-and-other-biologic.
46. Neuhäuser M, Thielmann M, Ruxton GD. The number of strata in propensity score stratification for a binary outcome. Archives of Medical Science. 2016.

IX. APPENDICES

A. SUMMARY OF PREVIOUS SURVIVAL TASK ORDER FINDINGS

A previous Mini-Sentinel workgroup (Survival Workgroup I; Task Order PI: Cook(44)) developed and began to evaluate new statistical methods to sequentially monitor rare event outcomes that allow for chronically used exposures (e.g., drugs) and events that may occur distant in time from the initiation of drug use (e.g., acute myocardial infarction) and require survival techniques. The previous survival workgroup concentrated their efforts on approaches using Cox's Proportional Hazards (PH) models(3) with direct adjustment for confounders in the regression model. They focused on methods that would be viable in the distributed data setting (e.g. subject-level data remains at the healthcare site behind firewalls and only deidentified data is shared across sites). Barriers to effective data sharing, such as privacy concerns and proprietary information policies, make pooling of subject-level data across sites rarely used unless deemed critical to the question of interest. They compared two approaches: 1) Cox PH regression adjusting for categorical confounders and aggregating survival/censoring times (deidentified Cox PH regression) and 2) Mantel Haenszel (MH) type estimate in which a Cox PH regression model is fit at each site and then the site-specific HR are pooled together using MH methods. They showed via a brief simulation evaluation that in this setting (distributed and rare event) and given a small number of confounders the new approaches were viable based on holding the overall type I error and minimizing bias. They compared the new approaches to standard methods not tailored to the Sentinel setting: 1) Cox PH regression directly adjusting for continuous confounders (non-distributed Cox PH regression) and 2) Site-stratified Cox PH regression in which one stratifies on site while still adjusting for other confounders in the model. The new approaches used permutation for boundary formation which was shown to be necessary if either the outcome and/or exposure were relatively rare. Below we briefly describe the main findings and limitations.

MAIN FINDINGS:

Under a more common outcome rate (0.05 per 10,000 study sample size yields ~500 total events), they found that methods perform well across all scenarios studied, and that the distributed data methods had similar power as the non-distributed data setting.

- Under a more rare outcome rate (0.01 per 10,000 study sample size yields ~100 total events), they found that for the more common exposure rate (proportion exposed = 0.50), the Cox PH regression methods (both non-distributed and deidentified) outperformed the Site-Stratified Cox PH regression and MH type estimates. Under less common exposure (proportion = 0.10), all methods did not perform as well indicated by elevated type I error especially the MH method, but also Cox PH regression adjusting directly for confounders. They then used permutation statistical inference and found methods performed better but still some elevated type I error for the distributed data approaches. Further, the MH indicated some loss of power.

LIMITATIONS:

- The previous work applied a very simple confounder model, adjusting only for two sites and age. Age when categorized had 5 indicator variables. There is a need to conduct a study using closer to real data that would be observed in Sentinel.
- In scenarios with more confounders, there is a need to explore methods such as propensity scores that reduce the number of parameters to include in the model.

- The MH method did not perform well, so we need to assess other approaches to stratification such as propensity scores. Approaches to applying these methods in the Sentinel context needs to be explored.

B. APPENDIX TABLES AND FIGURES FOR SECTION II

Table B 1. Pearson correlation matrix for binary and categorical variables (n=150,000)

Site 1						
	HS	1+ EV	1+ CS	SexF	Age Cat	Year
1+ HS	1	0.436	0.331	0.014	0.077	-0.013
1+ EV	0.436	1	0.277	0.02	0.007	0.007
1+ CS	0.331	0.277	1	0.027	0.14	0.005
SexF	0.014	0.02	0.027	1	-0.014	-0.007
Age Cat	0.077	0.007	0.14	-0.014	1	0.005
Year	-0.013	0.007	0.005	-0.007	0.005	1
Site 2						
	HS	1+ EV	1+ CS	SexF	Age Cat	Year
1+ HS	1	0.462	0.367	0.007	0.098	0.012
1+ EV	0.462	1	0.318	0.016	0.058	0.045
1+ CS	0.367	0.318	1	0.034	0.165	0.059
SexF	0.007	0.016	0.034	1	0.029	0.011
Age Cat	0.098	0.058	0.165	0.029	1	0.021
Year	0.012	0.045	0.059	0.011	0.021	1
Site 3						
	HS	1+ EV	1+ CS	SexF	Age Cat	Year
1+ HS	1	0.17	0.378	-0.016	0.096	0.008
1+ EV	0.17	1	0.167	0.031	-0.027	0.029
1+ CS	0.378	0.167	1	-0.014	0.181	0.038
SexF	-0.016	0.031	-0.014	1	0.022	0.008
Age Cat	0.096	-0.027	0.181	0.022	1	0.048
Year	0.008	0.029	0.038	0.008	0.048	1
Site 4						
	HS	1+ EV	1+ CS	SexF	Age Cat	Year
1+ HS	1	0.091	0.394	0.018	0.114	0.003
1+ EV	0.091	1	0.123	0.035	-0.056	0.011
1+ CS	0.394	0.123	1	0.038	0.181	0.012
SexF	0.018	0.035	0.038	1	0.053	-0.004
Age Cat	0.114	-0.056	0.181	0.053	1	0.031
Year	0.003	0.011	0.012	-0.004	0.031	1
Site 5						
	HS	1+ EV	1+ CS	SexF	Age Cat	Year
1+ HS	1	0.444	0.341	0.025	0.023	0.012
1+ EV	0.444	1	0.281	0.023	-0.016	0.022
1+ CS	0.341	0.281	1	0.038	0.072	-0.005
SexF	0.025	0.023	0.038	1	-0.015	0.022
Age Cat	0.023	-0.016	0.072	-0.015	1	0.067
Year	0.012	0.022	-0.005	0.022	0.067	1

Table B 2. Site-specific regression chain coefficients for binary covariates (n=150,000)

Site 1				
	Int	1+ HS	1+ EV	1+ CS
1+ HS	0.101			
1+ EV	-1.859	2.734		
1+ CS	-1.779	1.593	0.942	
SexF	-0.069	0.008	0.064	0.114
Site 2				
	Int	1+ HS	1+ EV	1+ CS
1+ HS	0.097			
1+ EV	-1.884	2.964		
1+ CS	-1.740	1.778	1.085	
SexF	0.007	-0.072	0.050	0.167
Site 3				
	Int	1+ HS	1+ EV	1+ CS
1+ HS	0.160			
1+ EV	-1.905	1.045		
1+ CS	-1.036	2.100	0.685	
SexF	0.060	-0.092	0.198	-0.055
Site 4				
	Int	1+ HS	1+ EV	1+ CS
1+ HS	0.144			
1+ EV	-2.028	0.663		
1+ CS	-1.649	2.167	0.643	
SexF	-0.106	0.009	0.185	0.157
Site 5				
	Int	1+ HS	1+ EV	1+ CS
1+ HS	0.107			
1+ EV	-1.742	2.765		
1+ CS	-1.697	1.626	0.903	
SexF	-0.090	0.066	0.046	0.151

Abbreviations: HS=Hospital Stay, EV=Emergency Department Visit, CS=Comorbidity Score, SexF=Female

Table B 3. Site-specific regression chain coefficients for categorical variables (n=150,000)

Site 1		Int	1+ HS	1+ EV	1+ CS	SexF	45-54	55-64	≥65
Age	45-54	0.069	-0.175	-0.129	0.065	-0.055			
	55-64	0.009	0.111	-0.351	0.283	-0.114			
	≥65	-0.729	0.476	-0.345	0.997	-0.072			
Year	2009	0.719	-0.075	-0.033	-0.015	0.004	0.011	0.035	-0.010
	2010	0.557	-0.115	-0.008	0.000	-0.019	0.024	0.036	0.047
	2011	0.412	-0.186	0.003	0.007	-0.011	-0.070	-0.002	-0.042
	2012	0.296	-0.229	0.109	0.054	-0.043	-0.027	0.062	0.046
Site 2		**Int**	**1+ HS**	**1+ EV**	**1+ CS**	**SexF**	**45-54**	**55-64**	**≥65**
Age	45-54	0.154	-0.443	-0.080	0.090	-0.076			
	55-64	0.136	-0.171	-0.205	0.252	-0.045			
	≥65	-0.082	0.282	-0.074	0.937	0.128			
Year	2009	-0.164	-0.113	0.082	0.114	-0.044	-0.007	0.040	-0.022
	2010	-0.412	-0.290	0.205	0.198	0.023	0.016	0.035	0.052
	2011	-0.636	-0.176	0.184	0.246	0.084	-0.014	0.049	0.091
	2012	-0.648	-0.288	0.297	0.385	-0.010	-0.099	-0.024	0.028
Site 3		**Int**	**1+ HS**	**1+ EV**	**1+ CS**	**SexF**	**45-54**	**55-64**	**≥65**
Age	45-54	0.227	-0.160	-0.163	0.307	-0.056			
	55-64	0.371	0.022	-0.244	0.656	-0.007			
	≥65	1.545	0.255	-0.530	1.149	0.111			
Year	2009	-0.067	-0.068	0.057	0.103	-0.038	0.007	0.060	0.043
	2010	-0.263	-0.050	0.069	0.143	0.032	0.059	0.096	0.106
	2011	-0.482	-0.101	0.183	0.172	0.025	0.103	0.201	0.339
	2012	-0.394	-0.106	0.197	0.188	0.015	-0.048	0.193	0.259
Site 4		**Int**	**1+ HS**	**1+ EV**	**1+ CS**	**SexF**	**45-54**	**55-64**	**≥65**
	45-54	0.232	-0.046	-0.457	-0.029	-0.114			
Age	55-64	0.219	-0.020	-0.730	0.212	0.006			
	≥65	-0.137	0.398	-0.658	1.001	0.252			
	2009	-0.153	0.002	-0.002	0.003	0.036	0.118	0.132	-0.015
Year	2010	-0.248	-0.045	0.031	0.021	0.033	0.050	0.108	0.013
	2011	-0.310	-0.049	0.049	0.020	0.011	0.013	0.047	0.147
	2012	-0.389	-0.028	0.093	0.032	-0.047	-0.058	0.073	0.194
Site 5		**Int**	**1+ HS**	**1+ EV**	**1+ CS**	**SexF**	**45-54**	**55-64**	**≥65**
Age	45-54	0.324	-0.521	0.086	-0.144	-0.192			
	55-64	0.311	-0.306	-0.148	0.133	-0.104			
	≥65	-0.750	0.266	-0.331	0.612	-0.139			
Year	2009	0.099	0.196	-0.162	-0.031	-0.059	-0.124	0.013	0.149
	2010	-0.278	-0.265	0.017	-0.065	0.158	0.022	0.177	0.523
	2011	-0.200	0.102	-0.001	-0.269	0.079	0.068	0.105	0.385
	2012	-0.404	0.070	0.158	-0.072	0.086	-0.061	0.200	0.773

Figure B 1 a. Simulation distribution of propensity score coefficients from Site 1

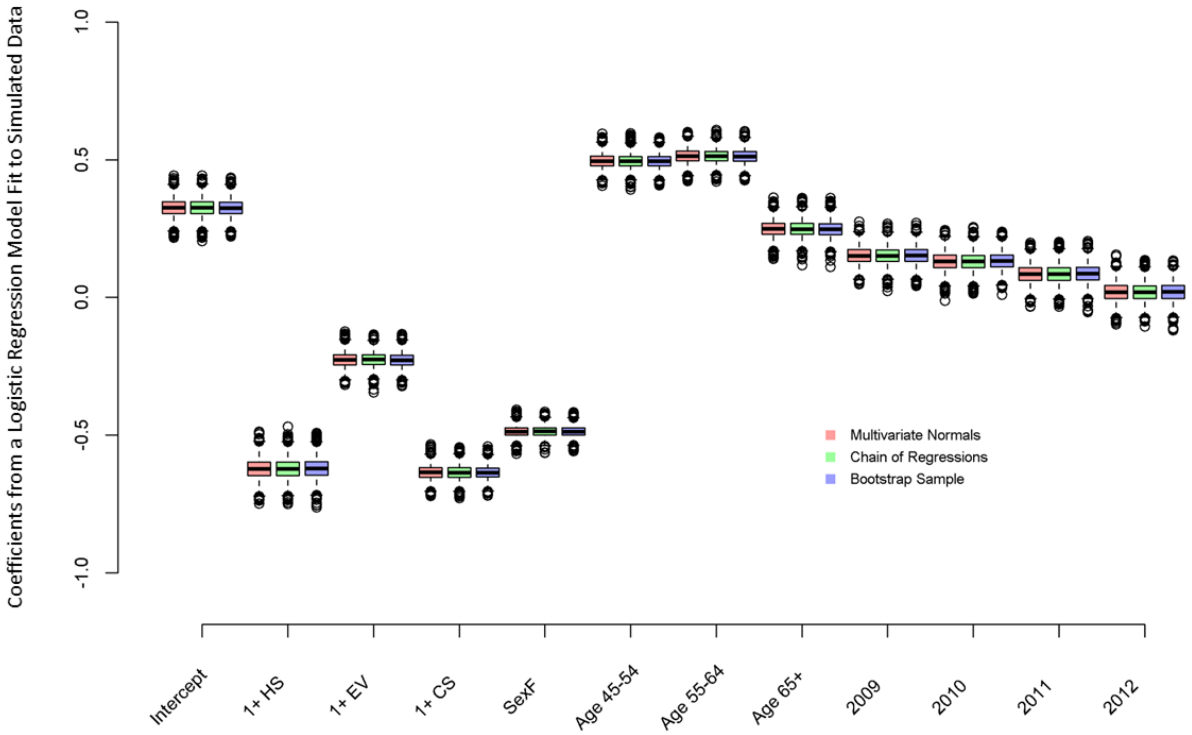


Figure B 1 b. Simulation distribution of propensity score coefficients from Site 2

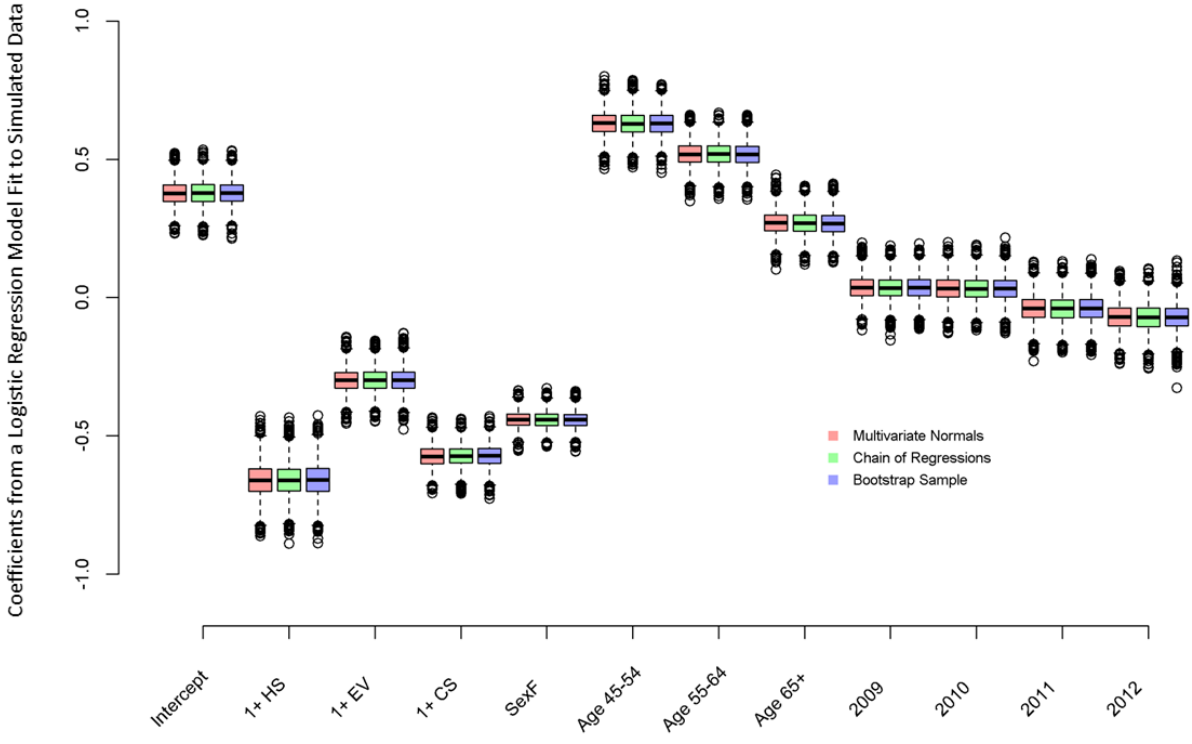


Figure B 1 c. Simulation distribution of propensity score coefficients from Site 3

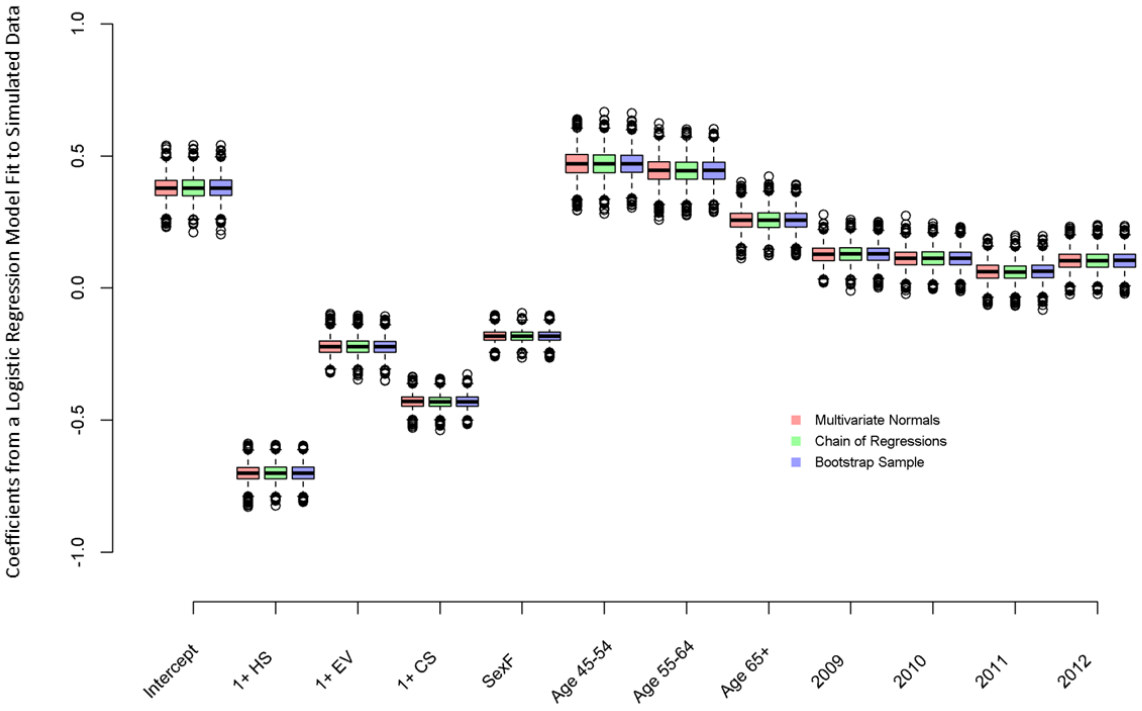


Figure B 1 d. Simulation distribution of propensity score coefficients from Site 4

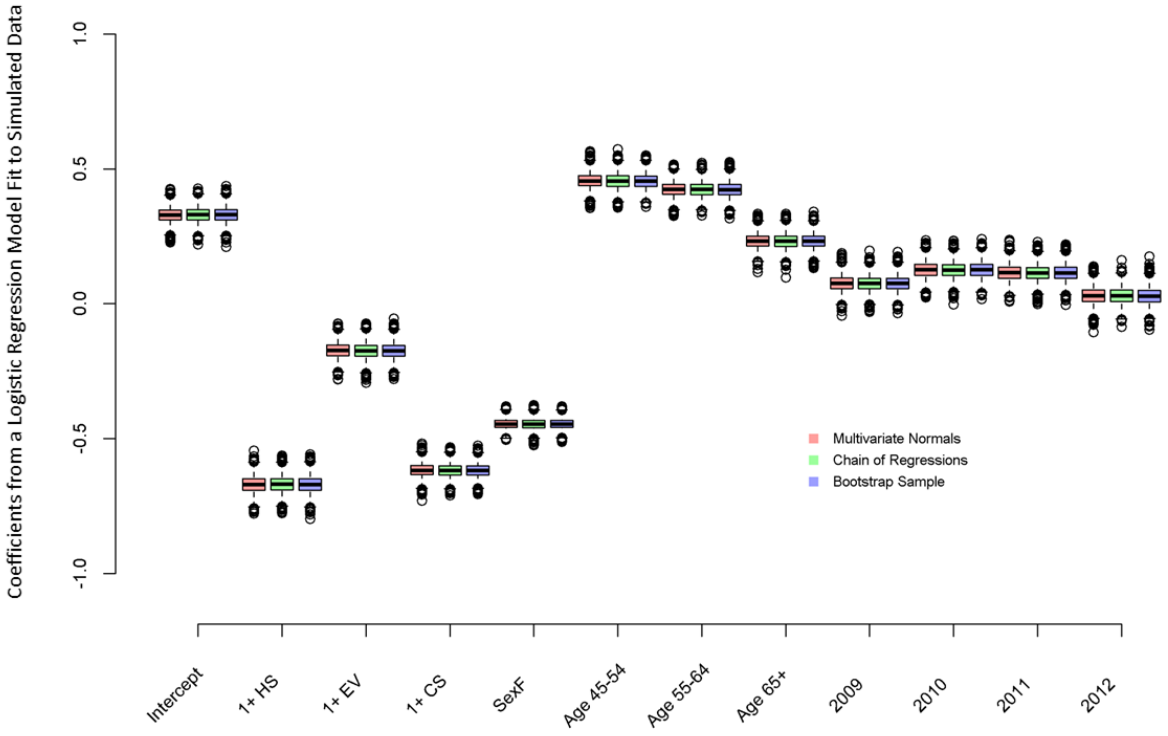


Figure B 1 e. Simulation distribution of propensity score coefficients from Site 5

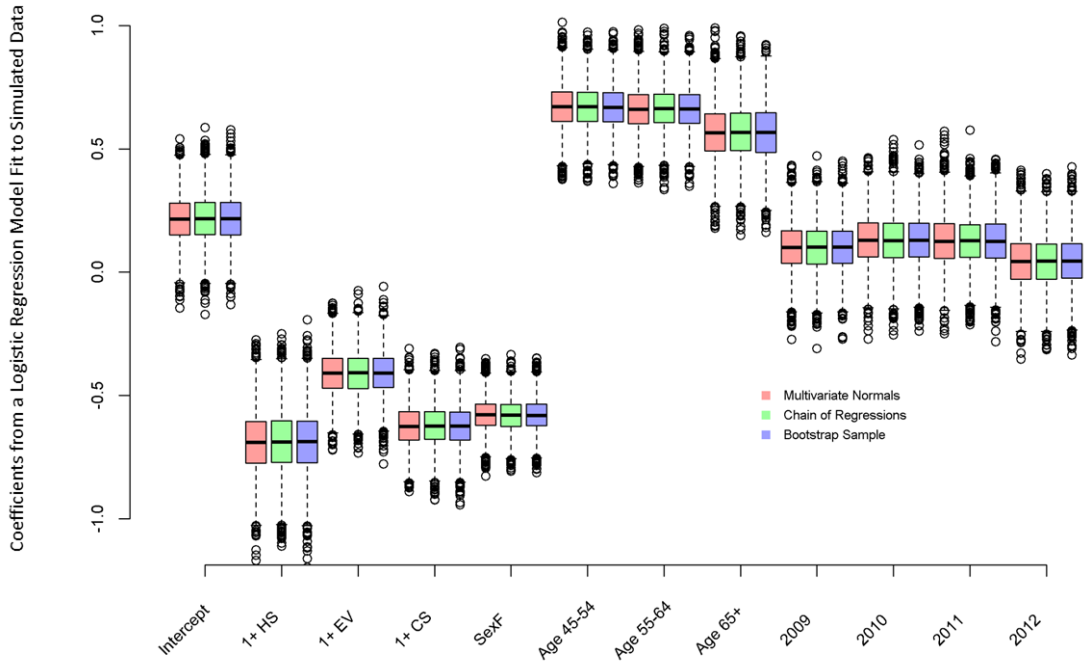


Figure B 2 a. Site 1 simulation distributions of coefficients from Cox PH outcome model with simple censoring (5,000 simulations)

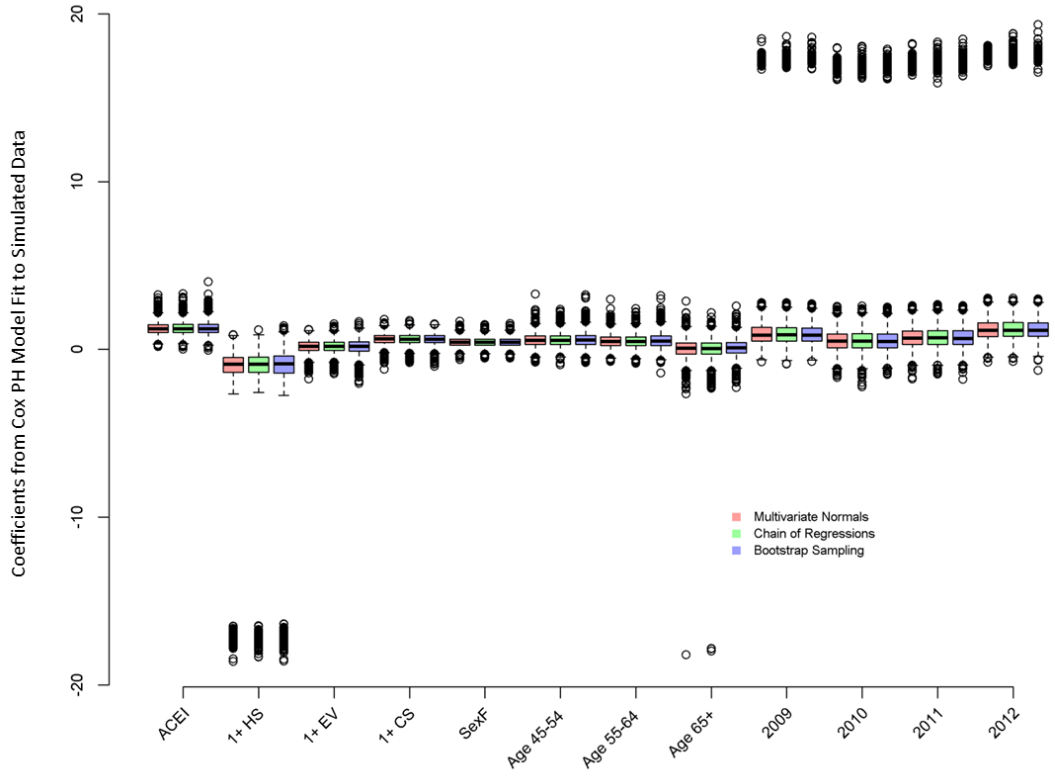


Figure B 2 b. Site 2 simulation distributions of coefficients from Cox PH outcome model with simple censoring (5,000 simulations)

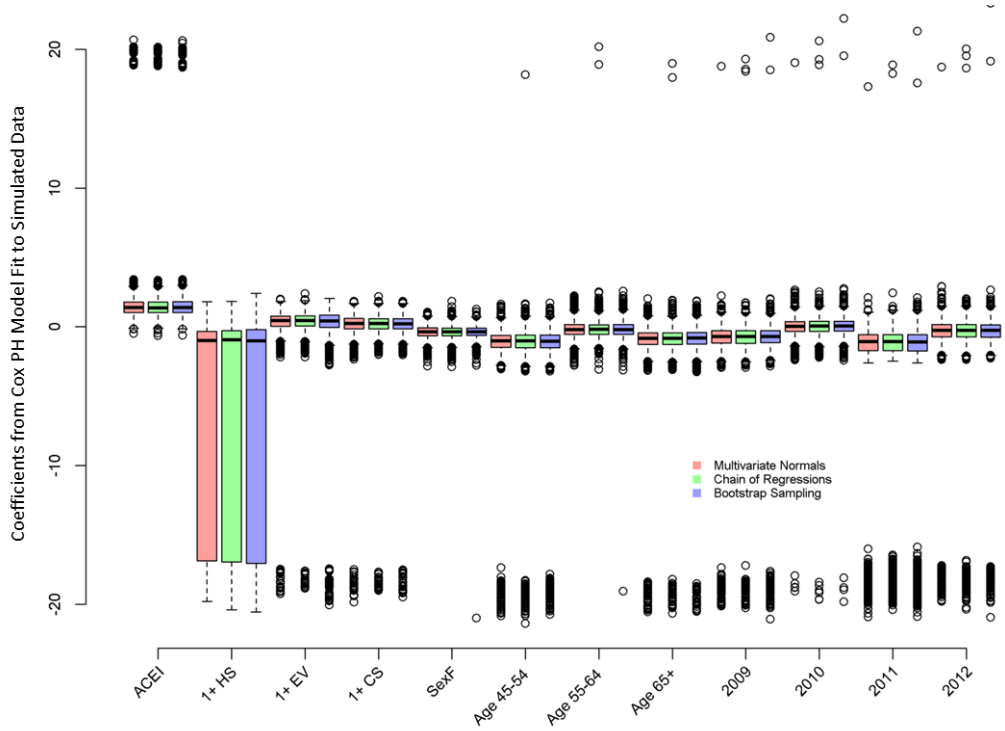


Figure B 2 c. Site 3 simulation distributions of coefficients from Cox PH outcome model with simple censoring (5,000 simulations)

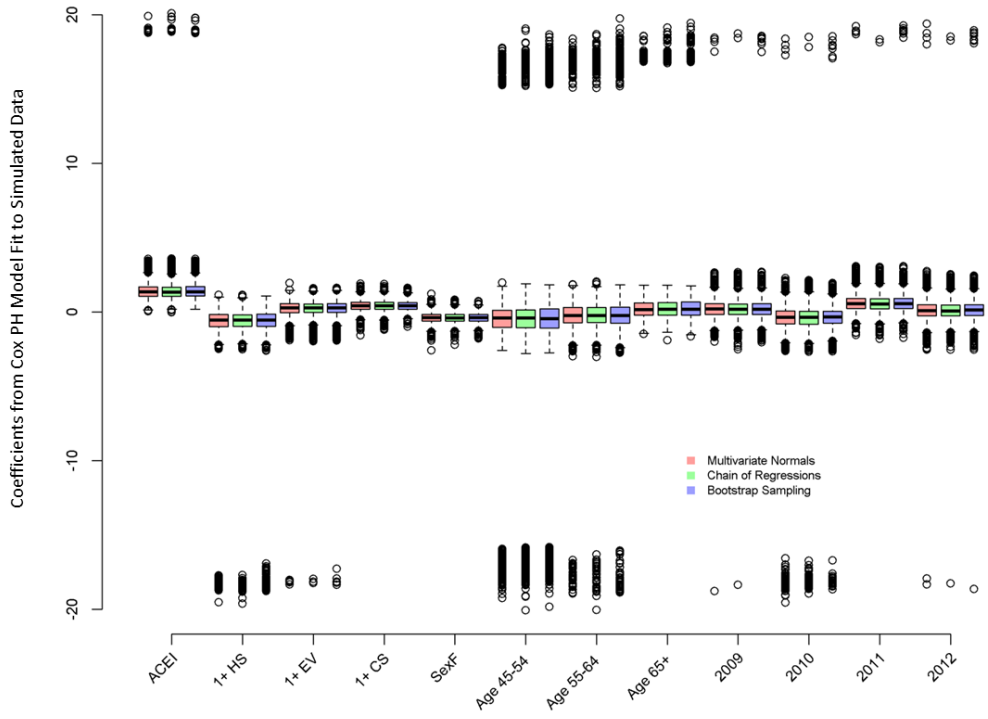


Figure B 2 d. Site 4 simulation distributions of coefficients from Cox PH outcome model with simple censoring (5,000 simulations)

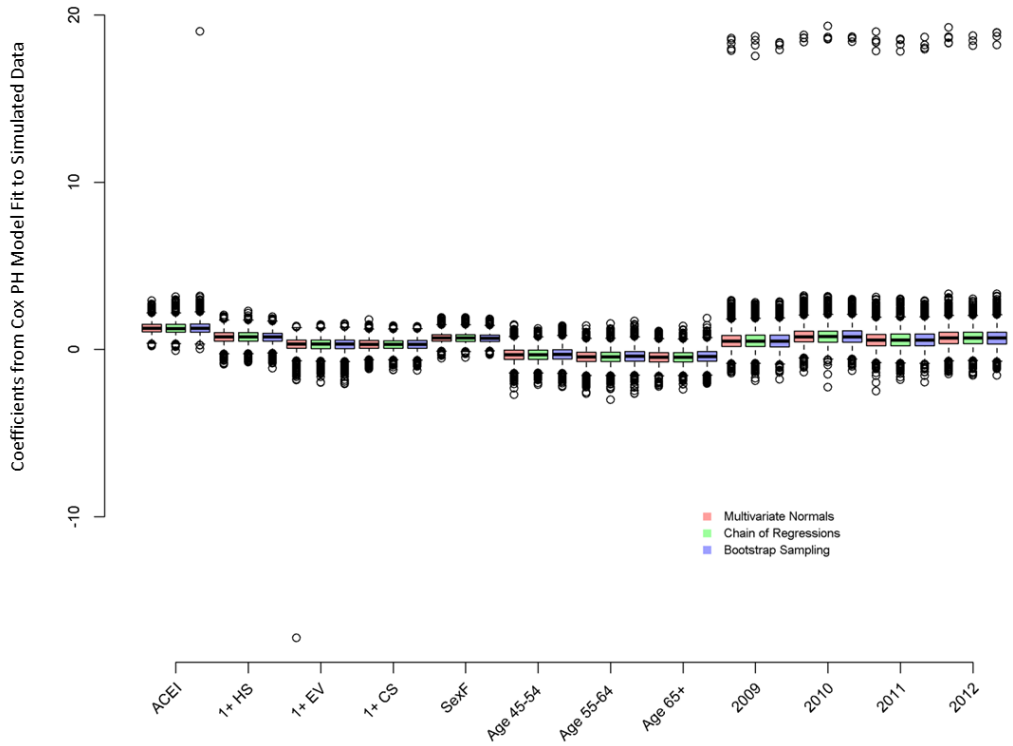


Figure B 2 e. Site 5 simulation distributions of coefficients from Cox PH outcome model with simple censoring (5,000 simulations)

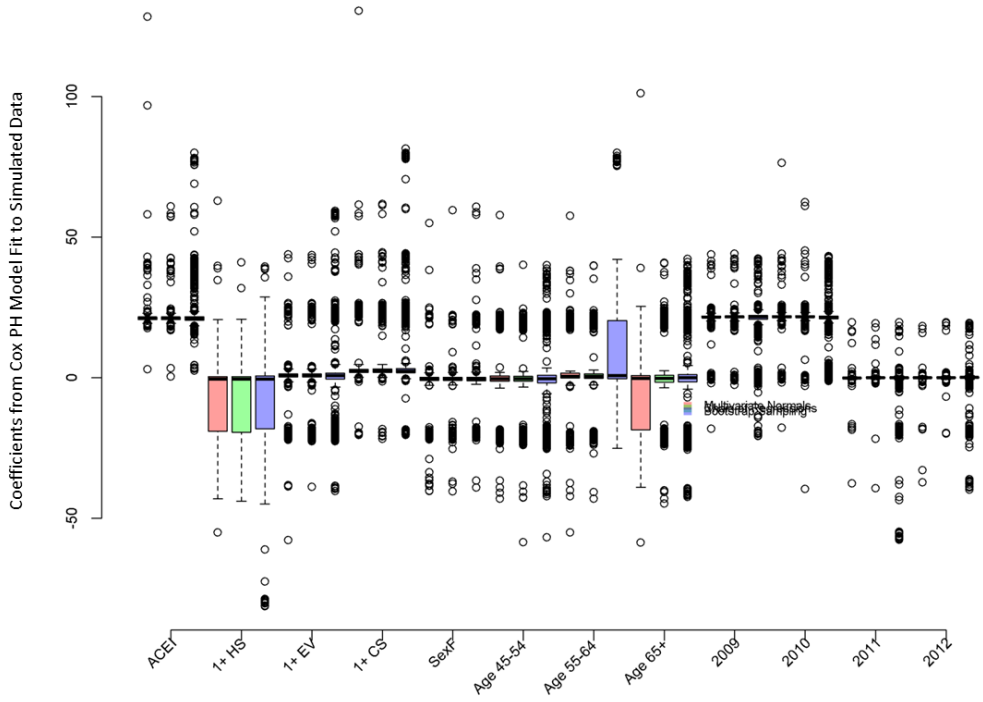


Figure B 3 a. Site 1 simulation distributions of coefficients from Cox PH outcome model with simple censoring with points (5,000 simulations)

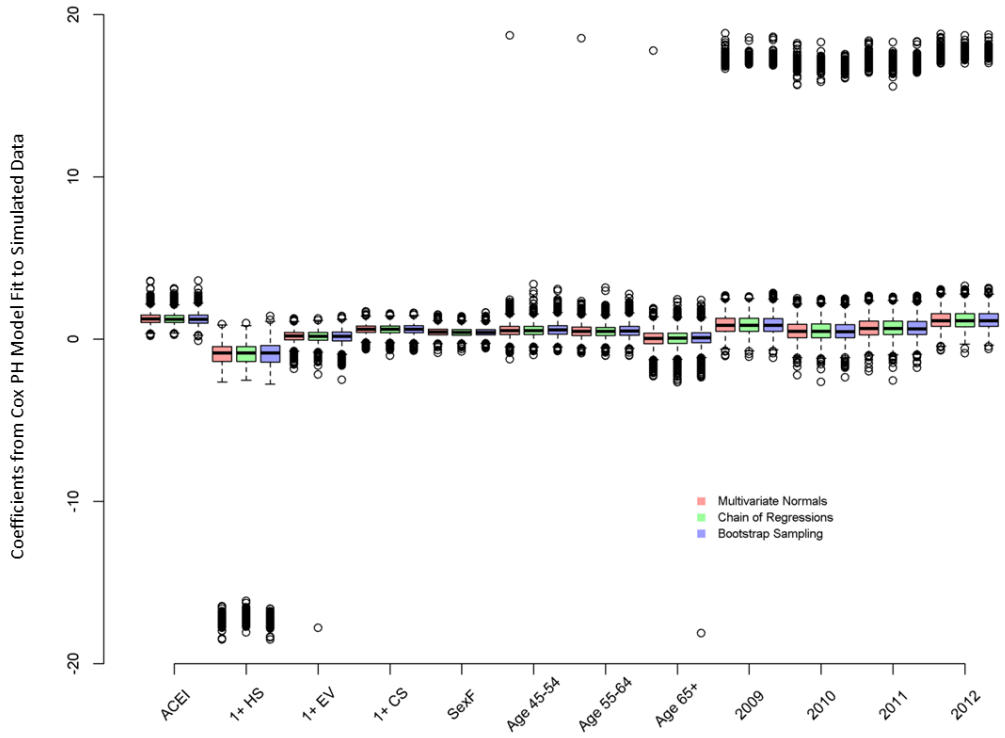


Figure B 3 b. Site 2 simulation distributions of coefficients from Cox PH outcome model with simple censoring with points (5,000 simulations)

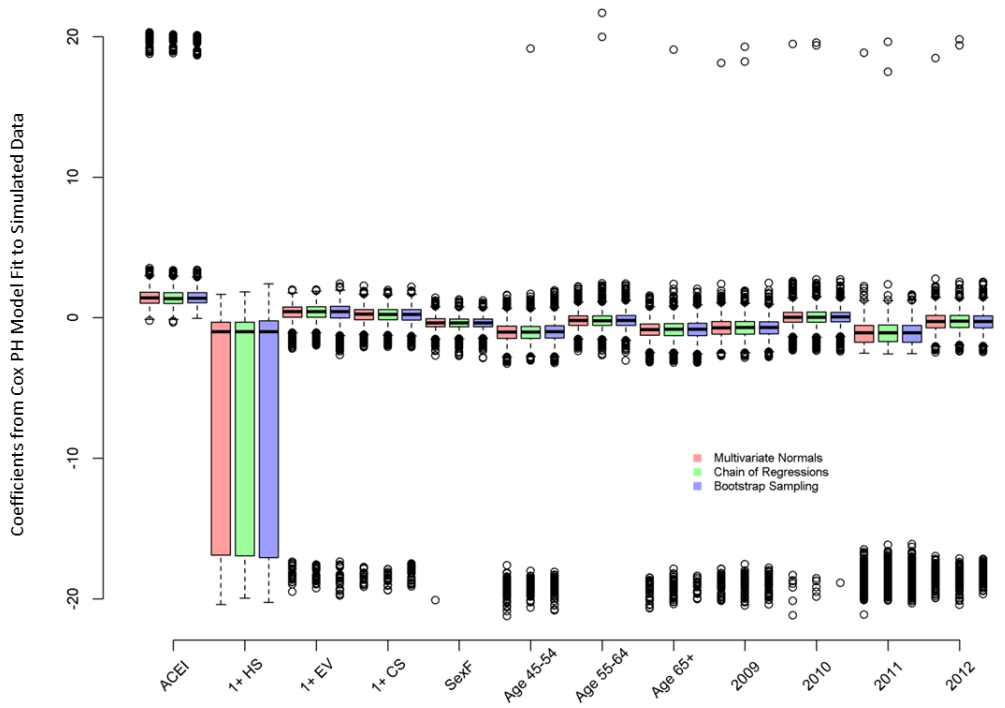


Figure B 3 c. Site 3 simulation distributions of coefficients from Cox PH outcome model with simple censoring with points (5,000 simulations)

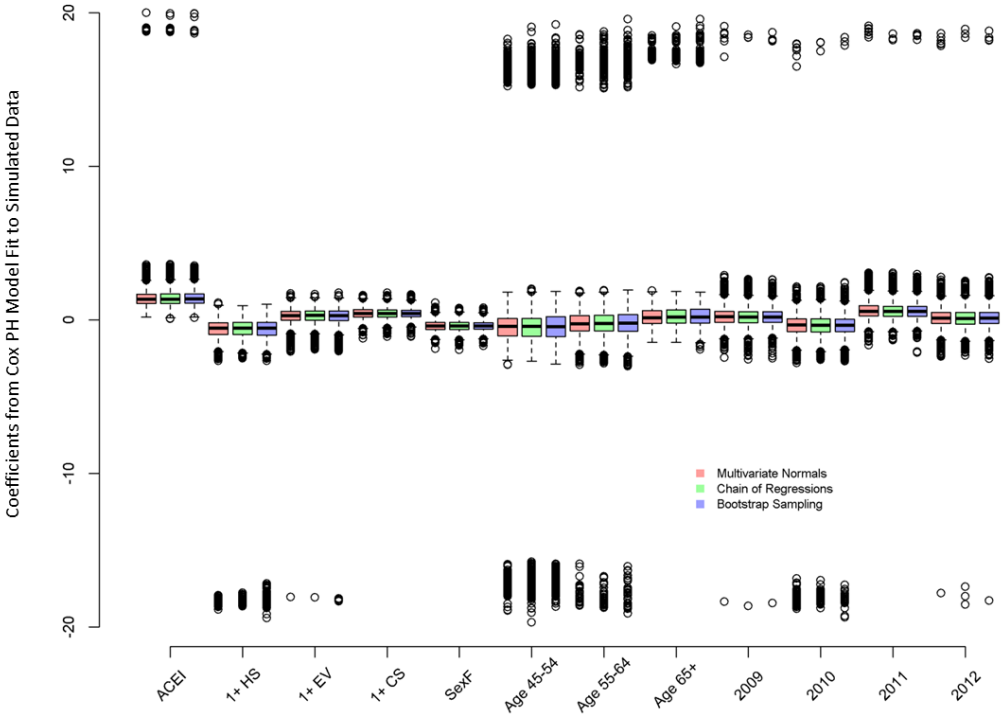


Figure B 3 d. Site 4 simulation distributions of coefficients from Cox PH outcome model with simple censoring with points (5,000 simulations)

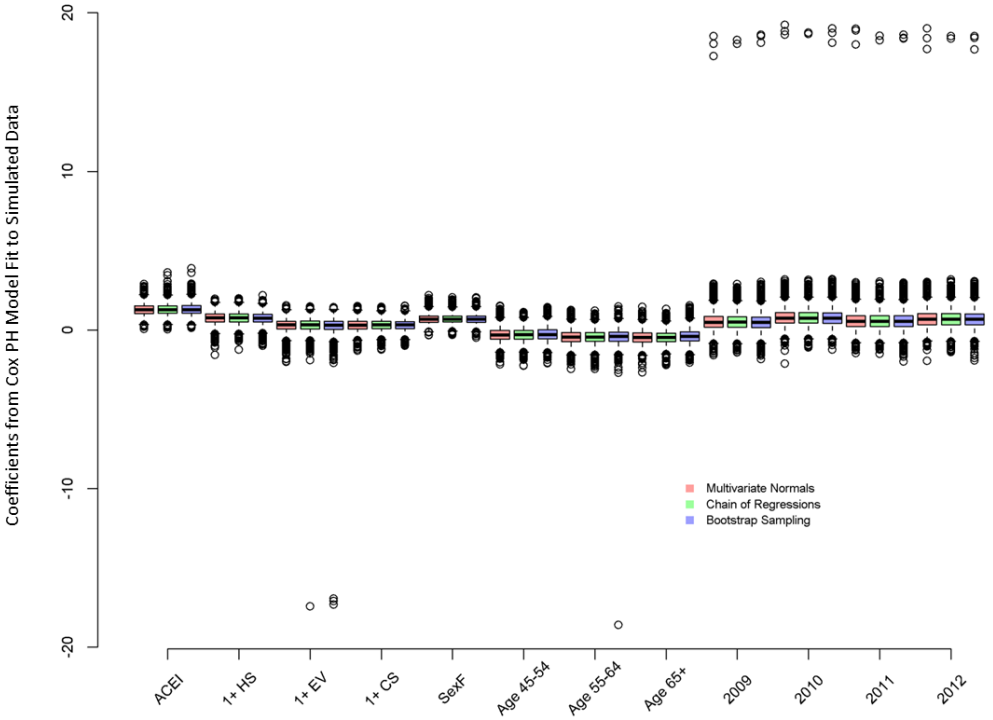


Figure B 3 e. Site 5 simulation distributions of coefficients from Cox PH outcome model with simple censoring with points (5,000 simulations)

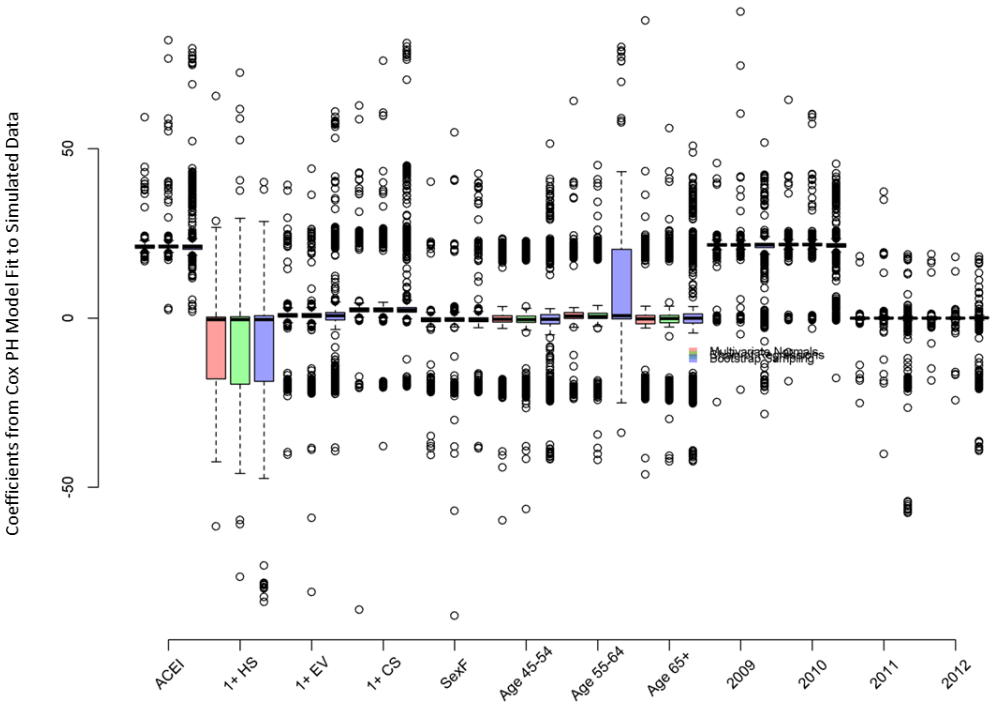


Figure B 4 a. Site 1 simulation distributions of coefficients from Cox PH outcome model with covariate adjusted censoring (5,000 simulations)

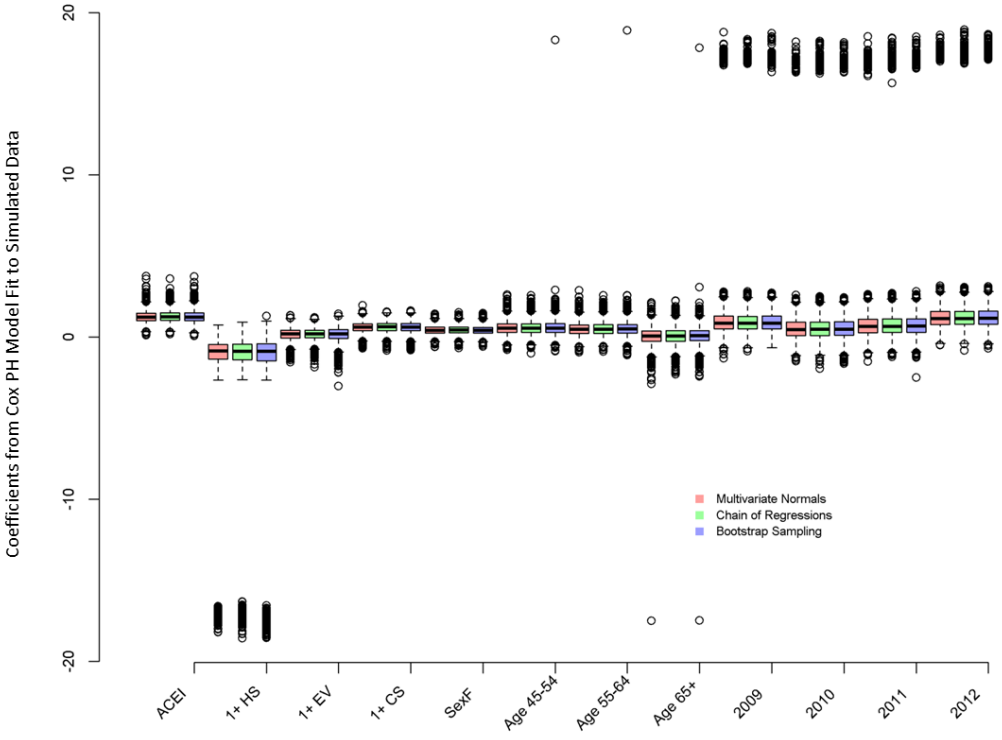


Figure B 4 b. Site 2 simulation distributions of coefficients from Cox PH outcome model with covariate adjusted censoring (5,000 simulations)

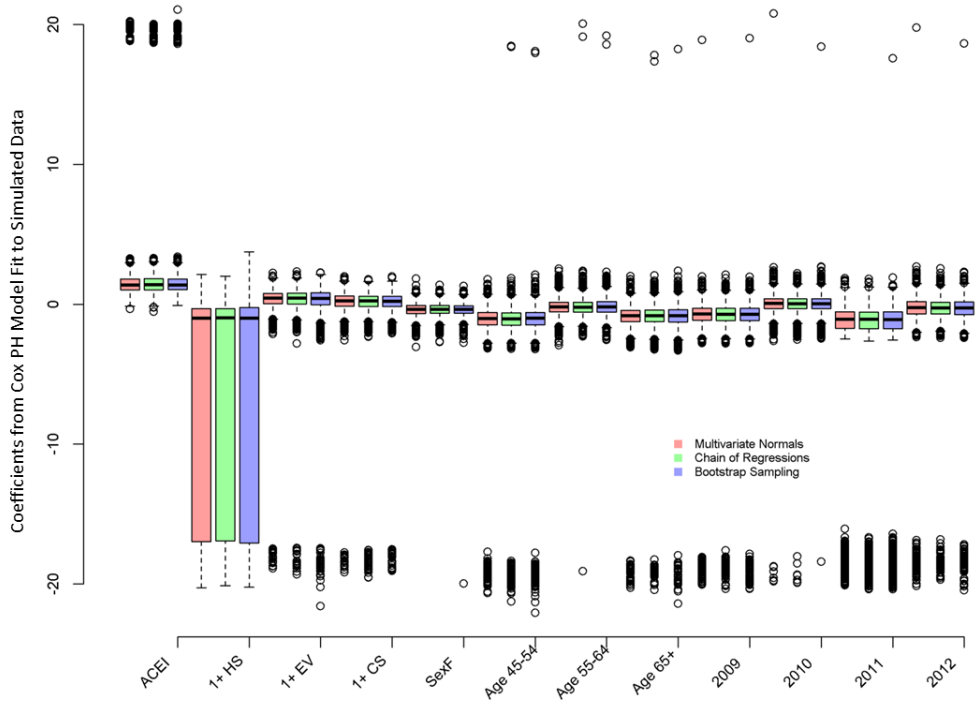


Figure B 4 c. Site 3 simulation distributions of coefficients from Cox PH outcome model with covariate adjusted censoring (5,000 simulations)

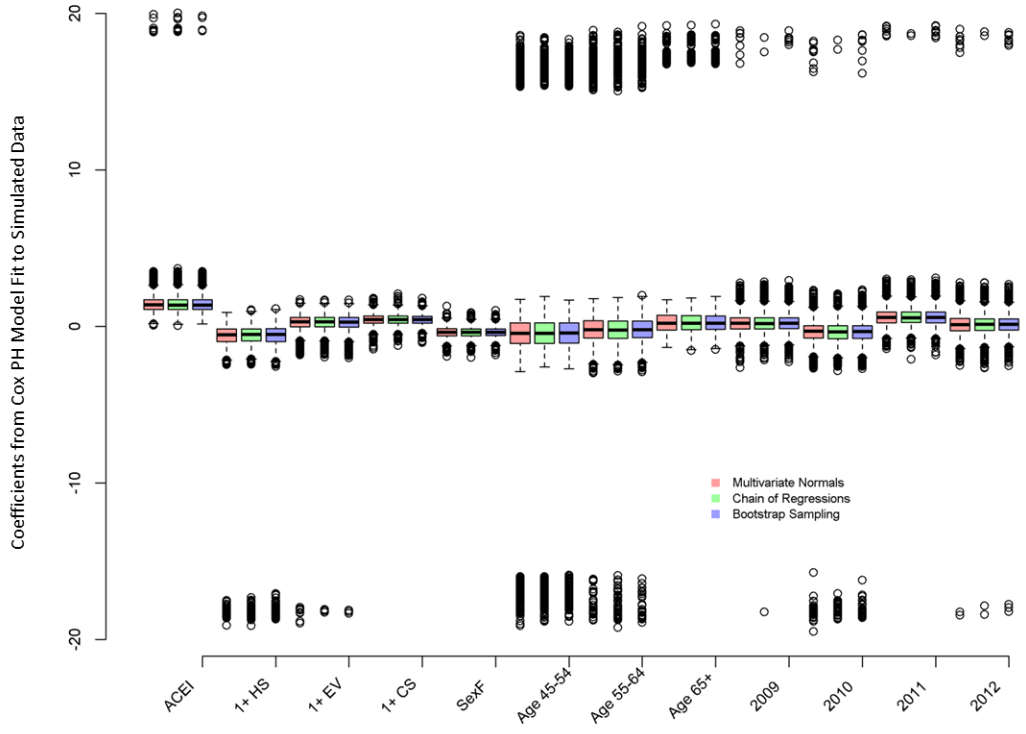


Figure B 4 d. Site 4 simulation distributions of coefficients from Cox PH outcome model with covariate adjusted censoring (5,000 simulations)

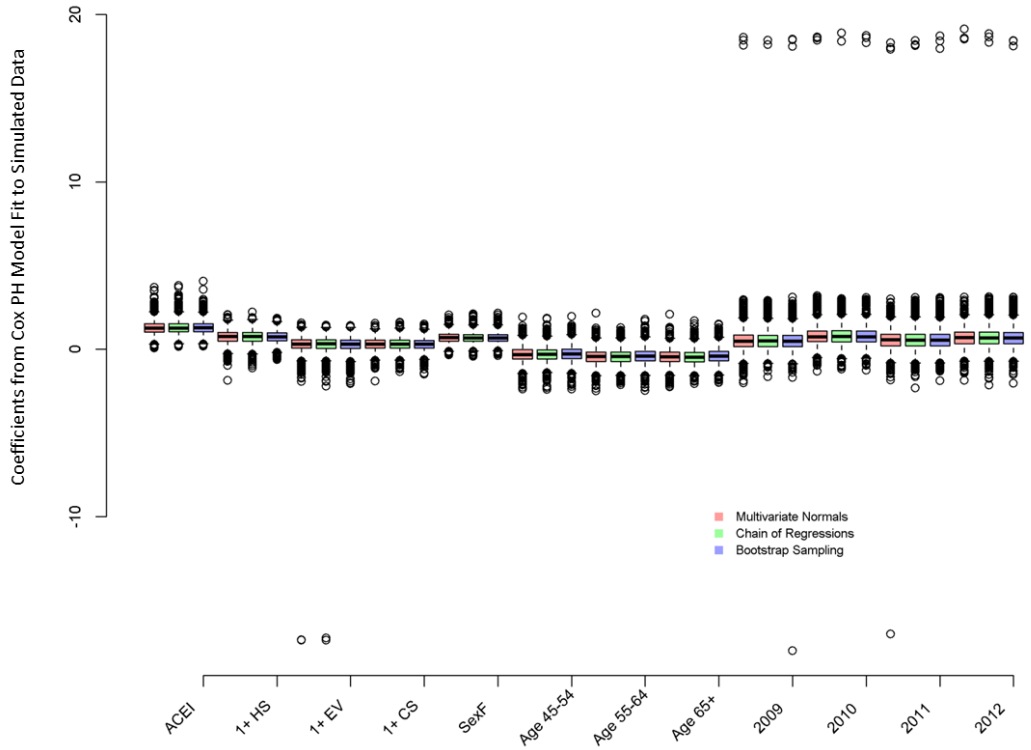
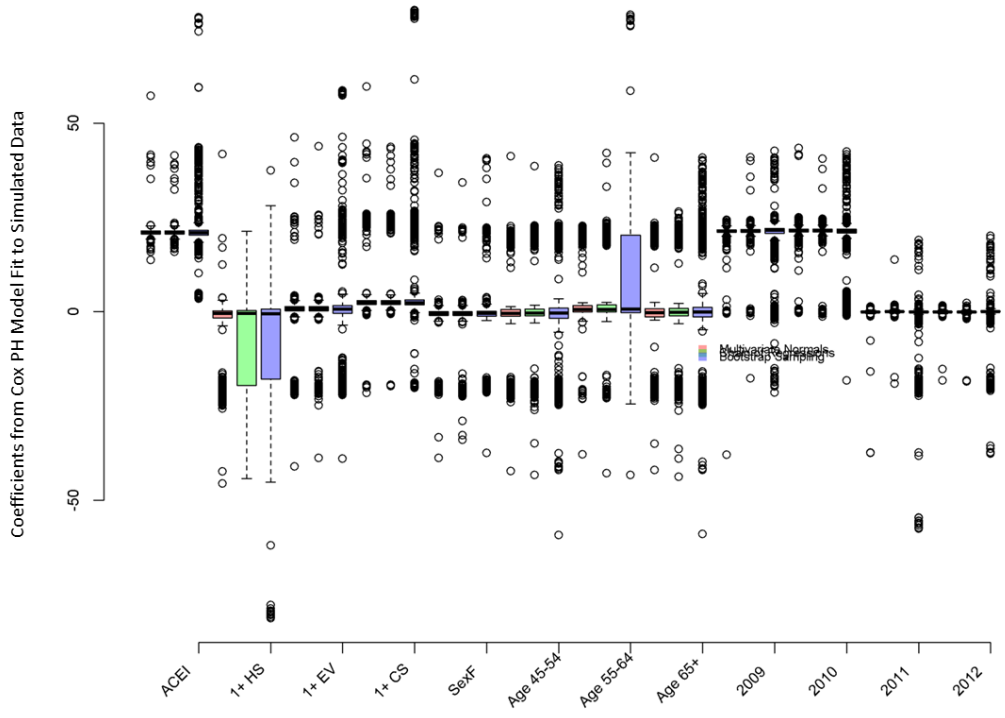


Figure B 4 e. Site 5 simulation distributions of coefficients from Cox PH outcome model with covariate adjusted censoring (5,000 simulations)



C. SAFETY SURVEILLANCE AND THE ESTIMATION OF RISK IN SELECT POPULATIONS: FLEXIBLE METHODS TO CONTROL FOR CONFOUNDING WHILE TARGETING MARGINAL COMPARISONS

1. Introduction

Electronic health records (EHR) data have provided the opportunity for new research to improve public health. An important national effort is the Food and Drug Administration (FDA) Sentinel initiative program which has created a surveillance network with over 100 million patient lives to monitor the safety of approved medical products. One of the interests is to estimate the effect of exposure on the overall risk of binary adverse events in a select population with comparison on the population level, which may not be fully powered for nor be targeted at in randomized controlled trials. EMR data provides not only sufficient sample sizes but also extensive patient features recorded over time that allows robust and efficient inference.

Use of large scale administrative EMR data for drug safety research comes with challenges. A key challenge is the need to control for a large number of confounders, when drug adverse events are often rare. Regression adjustment of many confounders for rare outcomes may have model fitting issues. In contrast, the exposure is usually sufficient, and thus the probability of being exposed, i.e. the propensity score, can be predicted by the rich patient information. In such a situation, regression adjustment of the propensity score, a one-dimensional summary score known to be sufficient for balancing the exposure and control groups (Rosenbaum & Rubin 1983)(6), is advantageous.

In a regression model that estimates exposure effect controlling for propensity score, the adjusted coefficient of the exposure has a conditional interpretation, i.e., a comparison of the risks among restricted group of homogeneous patients having the same characteristics. In drug safety research, we are often more interested in a marginal effect and care about generalizability to the full population, which is a combination of the exposed and unexposed groups. Such an effect is a comparison of risks estimated using the full population containing heterogeneous patients.

To make population-level comparison, a common strategy used in epidemiology literature is direct standardization or direct adjustment. It applies stratum-specific rates observed in the exposed and unexposed groups to the full population in order to obtain the number of events expected in the full population under exposure and control, as well as estimate the population-level risks. Through this approach, one is able to control for a confounder while targeting a population-level comparison.

One drawback of direct standardization is that it applies to one single confounder. Regression allows us to control for multiple confounders that are either continuous or categorical. To make population-level comparison from a regression model, it has been proposed to take the empirical averages of the pair of predicted risks under exposure and unexposed of each subject. Such a procedure is called standardization,(29) but is also called G-computation,(9, 10) partial means,(19) marginal integration,(18, 20) full imputation,(17, 32) or marginalization(13, 25) in the literature.

From the above discussion, we see that adjustment of propensity score in a regression model followed by standardization to get back to the full population level is tailored to our particular question of interest in the specific post-marketing drug safety surveillance setting.

It has been used in previous literature in the past decade. Austin et al. (2007)(30) and Austin (2007)(31) compared propensity score methods and concluded that regression adjustment on the propensity score can result in biased exposure effect. However, one limitation of their papers was that the propensity

score was adjusted as a linear term on the probability scale, which may not capture the relationship between the outcome and the propensity. Thus, the biased exposure effect that was observed could be due to model misspecification rather than validity of propensity score adjustment. In addition, they took the regression coefficient for the treatment as the estimated "marginal" odds ratio (hazard ratio), which is in fact a conditional effect that adjusted for the propensity score. In general, we need both flexible methods for regression adjustment of propensity score, and comprehensive and valid simulation study to compare causal inference methods for binary outcomes.

In this paper, we propose adjustment of B-splines of the propensity score, which corrects for the bias from linear adjustment. We focus on binary outcomes, and use standardization to estimate the marginal, population-level mean of the potential outcomes. With the estimated mean outcomes (mean risks), one can obtain parameters of interest that have causal interpretation, such as risk difference, risk ratio, and odds ratio. **Section IX.C.2** provides brief background in causal inference and introduces notation. In **Section IX.C.3** we introduce the regression adjustment on propensity score method in detail and provide an empirical estimator of the variance. **Section IX.C.4** overviews existing causal inference propensity score methods which estimate the exposure effect targeting a certain population. In **Section IX.C.5**, we conduct simulation study to compare the flexible regression adjustment of propensity score with existing causal inference methods. We provide discussion and future work in Section **Section IX.C.6**.

2. Background in Causal Inference

a. The Potential Outcomes Framework

Causation is inferred by any observed difference between the mean outcomes under exposure and control holding everything else the same. Accordingly, for each subject i, we define a pair of estimands $(\mathrm{Y_i}(1),\mathrm{Y_i}(0))$ as the outcomes that would be observed under exposure and control, called the potential outcomes. Denote the binary exposure as $\mathrm{X_i}, \mathrm{i} = 1, \ldots, \mathrm{n}$ for subject i, taking on value 1 (exposed) or 0 (unexposed). For each subject, only one of the potential outcomes is observed, i.e., the observed outcome $\mathrm{Y} = \mathrm{Y}(1)$ if exposed ($\mathrm{X} = 1$) with $\mathrm{Y}(0)$ missing, and $\mathrm{Y} = \mathrm{Y}(0)$ if unexposed ($\mathrm{X} = 0$) with $\mathrm{Y}(1)$ missing.

b. The Strongly Ignorable Treatment Assignment Assumption

The gold standard for estimating a causal effect is to conduct a randomized controlled experiment, in which the exposed and unexposed groups are balanced. In particular, the mean of observed outcome in the exposed group, $\mathrm{E}[\mathrm{Y}|\mathrm{X} = 1]$, will be equal to the mean of potential outcome under exposure in the entire population, $\mathrm{E}[\mathrm{Y}(1)]$. Thus, one can directly estimate the population average using the observed portion.

In observational studies, however, differences in the outcomes between the two arms could be due to both pre-existing systematic differences and the drug effect. In the presence of confounding effects in observational studies, Rosenbaum and Rubin (1983)(6) proposed the strongly ignorable exposure assignment assumption, which is $(\mathrm{Y}(1),\mathrm{Y}(0)) \perp \mathrm{X}|\mathbf{Z}$, where $\boldsymbol{Z}$ denotes the baseline covariates. It states that treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates. It allows one to estimate the within-strata average of potential outcomes using the observed portion, as if one conducted randomization within each stratum. That is, within a stratum of $\boldsymbol{Z}$, we have that exposed and unexposed groups are balanced, and thus the observed portion is representative of the entire stratum.

c. The Propensity Score

The propensity score is the probability of being exposed given the subject's characteristics, i.e. $S = P[Z = 1|X]$ (Rosenbaum & Rubin 1983(6)). It has two important roles. First, it is a summary score that reduces the dimension: it summarizes a vector of the baseline covariates according to how predictive they are for measuring exposure-proneness, into a scalar. Second, it is a balancing score: conditional on the propensity score, the baseline covariates are similar between exposure and control groups. In practice it is often estimated assuming a logistic regression model and therefore $\hat{S}_i = (1 + \exp(\hat{\gamma}\mathbf{Z_i}))^{-1}$

d. Causal Inference in Observation Study

Causal inference is a comparison of the population-level average of the potential outcomes. The most common form of comparison is the mean difference, i.e., the causal exposure effect is measured as the average treatment effect, ATE=$E[Y(1)] - E[Y(0)]$, or the average treatment effect on the treated, ATT=$E[Y(1)|X = 1] - E[Y(0)|X = 1]$.

With estimating the population average of potential outcomes as the ultimate goal, causal inference methods either provide a balanced population that mimics one from a randomized experiment, or impute the unobserved potential outcomes. A review of causal inference methods using the propensity score will be provided in **Section IX.C.4**.

3. Standardization Using Flexible Propensity Score Regression

In this section we propose a method that flexibly adjusts for confounding using a propensity score, but then is able to standardize to any marginal estimand of interest. This is similar to standardization with direct adjustment for confounders(9, 10) that will be discussed in **Section IX.C.4** except that we have further incorporated a propensity scores flexibly in the model to reduce the dimensionality of the confounder adjustment while attempting to minimize model assumptions. We further have derived variance estimates that incorporate the variability due to the estimation of the propensity score.

a. Generalized Partially Linear Model

We propose a generalized partially linear outcome model as follows,

$$g\left(E\left[Y_i|X_i, \hat{S}_i\right]\right) = \alpha\left(\hat{S}_i\right) + \beta X_i,$$

where $\hat{S} = \hat{P}[X|\mathbf{Z}]$ is the estimated propensity score, $\alpha(\cdot)$ is an unknown and potentially nonlinear function that adjusts for confounding effects, β is the conditional exposure effect, and $g(\cdot)$ is a link function. For binary outcomes $g(\cdot)$ is often the logit link function and the propensity score is also estimated using a logistic regression model.

To estimate the nonlinear function $\alpha(S)$, we apply a nonparametric regression technique, the polynomial spline regression.(5, 14) A spline is a piece-wise polynomial function that is smooth at the joint of each piece, called the knot. Any spline function on a given set of knots can be expressed as a linear combination of B-splines. Thus, we generate a set of B-spline basis functions, $\mathbf{B}(S) = [B_1(S), \ldots, B_K(S)]$, then fit the outcome on the basis functions and the exposure indicator. The potential outcomes under being exposed and unexposed for each subject i are thus predicted as

$$\hat{E}[Y_i(1)|S_i] = g^{-1}\left(\hat{\beta} + \boldsymbol{B}\left(\hat{S}_i\right)\hat{\boldsymbol{\alpha}}\right)$$

$$\hat{E}[Y_i(0)|S_i] = g^{-1}\left(\boldsymbol{B}\left(\hat{S}_i\right)\hat{\boldsymbol{\alpha}}\right)$$

The to obtain a population average effect we use a standardization/G-Computation approach to obtain the following population-level average of the potential outcomes,

$$\hat{p}_1 = \hat{E}[Y(1)] = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(\hat{\beta} + \boldsymbol{B}(\hat{S}_i)\hat{\boldsymbol{\alpha}}\right)$$

$$\hat{p}_0 = \hat{E}[Y(0)] = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(\boldsymbol{B}(\hat{S}_i)\hat{\boldsymbol{\alpha}}\right)$$

which will be used for causal comparison. For binary outcomes, we plug in such estimated mean risk to estimate the parameter of interest such as the risk difference, the relative risk, or the odds ratio.

b. Variance Estimation Incorporating Uncertainty of the Propensity Score

To derive the variance of this flexible standardized model there is a need to incorporate the variability of the propensity score, the variability due to flexibly regressing the propensity score onto the outcome, and variability due to the standardization step to the population-level estimand. Hahn & Ridder (2013)(38) studied inference for a general three-step estimator and derived an influence function that incorporates the uncertainty in each of the steps. We follow their proposed procedure and derived the following variance estimates.

The variance estimator for risk difference (RD) is

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\widehat{IF}_{1i} - \widehat{IF}_{0i}\right)^2,$$

for log risk ratio (RR) is

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{\widehat{IF}_{1i}}{\hat{p}_1} - \frac{\widehat{IF}_{0i}}{\hat{p}_0}\right)^2,$$

and for log odds ratio (OR) is

$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{\widehat{IF}_{1i}}{\hat{p}_1(1-\hat{p}_1)} - \frac{\widehat{IF}_{0i}}{\hat{p}_0(1-\hat{p}_0)}\right)^2,$$

where $\widehat{IF}_{1i} = \hat{E}[Y_i(1)|\boldsymbol{Z}_i] - \hat{p}_1 + \frac{X_i}{S_i}(Y_i - \hat{E}[Y_i(1)|\boldsymbol{Z}_i])$ and $\widehat{IF}_{0i} = \hat{E}[Y_i(0)|\boldsymbol{Z}_i] - \hat{p}_0 + \frac{1-X_i}{1-S_i}(Y_i - \hat{E}[Y_i(0)|\boldsymbol{Z}_i])$.

See Appendix for details of the derived variance estimation. When the outcome is rare, it could be difficult to directly estimate the variance and therefore bootstrap-based variance estimators may be needed. We will compare the performance of the empirical estimator and the bootstrapped variance estimator via simulation in **Section IX.C.5**.

4. Propensity Score Methods for Binary Outcomes

In this section, we review existing methods for estimating a population-level mean risk, which will be plugged in to estimate a population-level risk difference, risk ratio, or odds ratio. Recall that in **Section IX.C.2.d** we briefly introduced the main ideas in causal inference methods. As will be discussed in detail below, propensity score matching and propensity score stratification are two methods that mimics randomization to achieve balance in the two arms and thus make fair comparison; the inverse probability of treatment weighting reweights to a pseudo-population that is also balanced; the

regression based standardization method and the doubly robust estimator predicts the missing potential outcomes to make comparison using mean estimated from all subjects in the population.

a. Propensity Score Matching

The propensity score matching method mimics the randomized study by selecting a subpopulation, which includes matched sets of exposed and unexposed subjects sharing similar propensity scores. One common application of matching is to match each exposed participant to M unexposed participants. Then one would use regression to estimate the marginal causal estimand of interest. Specifically, if interest is in the marginal odds ratio one would fit a logistic regression model using data from the matched subpopulation and only include the indicator of exposed or unexposed in the model. Note that the matched subpopulation contains subjects with characteristics similar to the exposed arm. Thus, the estimated causal effect is in fact the average treatment effect on the treated (ATT).

In practice, applying propensity score matching involves several decisions to make. First of all, one needs to decide the value of M, as well as a caliper that defines the tolerance of the difference in propensity scores for a matched pair. It was discussed in simulation studies that increasing M tended to increase the bias but decrease the sampling variability of the estimate (Austin 2010)(34). The caliper can be decided in practice by checking the covariate balance as well as number of subjects in the matched dataset.

Second, one needs to choose a sampling method, i.e., with or without replacement. For matching without replacement, each unexposed subjected can be used at most once. For matching with replacement, a pseudo-population that is closest to the exposed population is generated. However, it is hard to interpret the result, and the possibility of including duplicated subjects needs to be accounted for when estimating the variance.

Third, matching has been implemented in several packages written in different statistical programming languages including R, SAS, and STATA. Each package has its own choice of algorithms and may therefore give different results. It is important to understand which algorithm is being used. Since the matching procedure does not involve the outcome, one could try multiple methods and select the best matched dataset according to covariate balance and size of the matched sample.

Last, for estimating the odds ratio, one could use the conditional logistic regression which fits the regression model acknowledging the fact that matched sets include similar subjects. However, the quantity being estimated becomes a conditional odds ratio, conditional on the matched set of similar subjects. The conditional logistic regression is implemented by applying the Cox proportional hazard model with tied survival times for subjects within the same matched set. There are different methods for dealing with tied survival times, depending on whether the likelihood function written in exact form or approximated form. The choice of methods affects the computation time and more importantly, the bias of the estimate, so sensitivity analysis on choice of methods for ties is recommended.

b. Stratification on Propensity Score

Propensity Score stratification, also referred to as subclassification, cuts the propensity score into strata according to its quantile, and then divides the population into equal-size subclasses of subjects having propensity score within the same strata (Rosenbaum & Rubin 1984)(7). Extreme subclasses with zero (un)exposed subjects will be non-informative and discarded, which is analogous to unmatched subjects in propensity score matching. Compared to matching, the stratification on propensity score also mimics randomization by achieving balance within subclasses. However, it includes more observations than

matching which reduces variance at the potential price of increased bias. The trade-off between bias and variance is controlled by the number of strata. An extreme case, for example, when the number of strata is equal to the number of matched pairs, some subclasses will have zero (un)exposed subjects and be discarded, and stratification and matching may result in similar estimates and datasets.

The common estimator following stratification is a weighted average of the quintile-specific odds ratios. For example, one weighting scheme leads to the Mantel-Haenszel estimator (Mantel & Haenszel 1959)(1). Adjusting for indicator of strata using a (conditional) logistic regression is another way to apply stratification for control of confounding. However, as mentioned in **Section IX.C.4.a**, conditional logistic regression will estimate a conditional odds ratio. Marginalization methods following regression will be introduced below in **Section IX.C.4.d**, which allows one to estimate a marginal causal odds ratio from a regression model.

c. Inverse Probability of Treatment Weighting (IPTW and Augmented IPTW)

Another way to achieve balance in the population is to reweight every subject to create a pseudo-population in which every (un)exposed pseudo-subject has equal possibility of being (un)exposed, which is representative to one from a randomized study. This is called the inverse probability of exposure weighting (IPTW).(26) A commonly used weight is the inverse of the propensity score, that is, to use $\frac{1}{S_i}$ if subject i is exposed and $\frac{1}{1-S_i}$ if subject i is unexposed. The idea behind using inverse probability is: for patients with a high S, they are more likely to be observed in the exposure group and more rarely seen in the control group, so using $\frac{1}{S}$ in the exposure group and $\frac{1}{1-S}$ in the control group will make the number of patients with the same value of S to be similar in the two groups. In other words, to achieve balanced groups of patients.

The idea of weighting is not new. In fact it has been widely used to achieve an estimator that is generalizable to a target population, especially in survey research.(21) To estimate the risk difference, one can take the weighted sum of the observed outcomes. To estimate the odds ratio or risk ratio, one can use a weighted generalized linear model to fit the outcome on exposure. Note that since the pseudo-population is balanced in terms of propensity score, and the only covariate in the model is the exposure, the estimated coefficient of the exposure is a marginal, causal effect.

A well-known problem with IPTW is the instability from inverting the estimated propensity score. Stabilized weights have been proposed (Robins et al. 2000)(26). A trimming approach, which truncates weight using either a pre-specified threshold or a quantile is also widely used in practice (Potter 1990, Potter 1993)(12, 16) and we will implement this in our simulation study in **Section IX.C.5**.

Simple IPTW requires that the propensity score must be correctly specified. To relax this assumption the Augmented IPTW (AIPTW) approach was proposed that developed a doubly robust estimator building on both the propensity score model and the outcome regression model (Robins et al. 1994, Bang & Robins 2005)(26, 27). It is doubly robust because it only requires either the propensity score model or the outcome model to be correctly specified. In addition, it was shown to be efficient among a class of semiparametric estimators, because it takes the efficient influence function as the estimating equation (Hahn 1998)(24). However, since the estimation of the outcome model is required there may be issues in the rare event setting relative to other methods. In the simulation approach we will only show the Augmented IPTW method since actual method performance between IPTW and Augmented IPTW was very comparable except in the case when the propensity score model was misspecified and Augmented IPTW performed better as expected.

d. Standardization with Direct Covariate Regression

As outlined in **Section IX.C.3** standardization after regression is a viable approach to estimate marginal estimands. We proposed the incorporation of a flexible propensity score model in the outcome regression model. However, the standard approach is to instead use a simpler two-step process which is also known as G-computation (Robins 1986, Robins 1987).(9, 10) The first step is to build an outcome regression model with both exposure and all confounders in the model

$$g(E[Y_i|X_i, \boldsymbol{Z}_i]) = \beta X_i + \boldsymbol{\alpha Z}_i.$$

Then standardization (Step 2) is done by simply taking the average of the predicted potential outcomes for all subjects in the population as outlined in **Section IX.C.3**. Then the estimated causal effect is a comparison of the marginal, population-level means, obtained by plugging in the marginalized mean into risk difference, risk ratio, or odds ratio.

Note that this is similar to the method proposed in **Section IX.C.3** except there is no estimation of the propensity score and the confounders are directly regressed on the outcome. Therefore, application of this method in the rare event setting may be problematic which will evaluate in the simulation study.

5. Simulation Study

We performed a simulation study to investigate the performance of the different methods outlined in **Section IX.C.3** and **4** to estimate a marginal OR. We chose to estimate a marginal OR since it is the most common estimand of interest in observational cohort studies. Our simulation study will mimic real data from a study comparing the effect of angiotensin-converting enzyme Inhibitors (ACEI) and beta blocker (BB) on incidence of angioedema in 30 days from the FDA Sentinel Initiative.

The marginal OR estimators to be compared are the following: (1) 1-1 matching on the propensity score without replacement; (2) Augmented IPTW, with parametric models for exposure and outcome, both adjusting for all covariates with trimmed propensity score using 5% tail as the threshold; (3) Standardization with direct covariate regression; (4) Standardization with regression on linear propensity score adjustment; (5) Standardization with regression on propensity score deciles; and (6) Standardization with flexible regression of the propensity score using B-spline basis functions (here we used cubic spline with one inner knot). We chose the first 3 methods since they are standard approaches used for estimation of a marginal OR. We chose approach 4 because this method has been shown to be biased in other simulation studies and therefore we were interested to assess for our simulation scenario if these findings still held. We chose approach 5 with standardization with regression on the propensity score deciles to be able to compare if the more flexible propensity score adjustment (approach 6) improved over this method.

Performance of approaches was assessed in terms of mean bias on log OR scale, type I error, and power. For each scenario assessed we used 8000 simulated datasets.

a. Simulation Setting

We generate a population of 100,000 subjects mimicking data from the ACEI and BB example. Specifically, there are nine binary clinically relevant covariates (NSAIDS, aspirin, ORAL-CS (optimizing recovery after laparoscopic colon surgery), allergic reaction, diabetes, heart disease, Ischemic HD, inpatient hospitalization, and gender) and one categorical variable which is age category with four levels, corresponding to three dummy variables (binary indicators). See **Table C 1Error! Reference source not found.** for prevalence of each confounder.

Table C 1. Prevalence of each confounder and relationship between exposure (ACEI and BB) and confounders for different simulation scenarios (propensity score model)

	Propensity Score Model Odds Ratios				
	%	Observed Propensity	Observed+ Age*Diabetes Interaction	Stronger Propensity	Stronger+ Age*Diabetes Interaction
Heart Disease	0.3	0.55	0.55	0.41	0.41
Aspirin	1.4	0.54	0.54	0.40	0.40
Ischemic HD	6.3	0.23	0.23	0.11	0.11
OptRec Colon Surg	8.0	0.85	0.85	0.78	0.78
Allergic Reaction	8.4	0.33	0.33	0.19	0.19
Inpatient Hosp.	10.5	0.86	0.86	0.80	0.80
NSAIDS	12.2	0.95	0.95	0.93	0.93
Diabetes	15.6	3.00	--	5.20	--
Female	47.3	0.51	0.51	0.36	0.36
Age(Ref: 18-44)					
45-54	32.5	1.49	--	1.82	--
55-64	27.6	1.37	--	1.60	--
65-99	7.4	1.08	--	1.12	--
Age*Diabetes (Ref: 18-44 and Not Diabetic)					
45-54 & Not Diabetic	27.4		1.49		1.82
55-64 & Not Diabetic	23.3		1.37		1.60
65-99 & Not Diabetic	6.2		1.08		1.12
18-44 & Diabetic	5.1		3.00		5.20
45-54 & Diabetic	5.1		7.37		15.58
55-64 & Diabetic	4.3		6.78		13.74
65-99 & Diabetic	1.2		5.34		9.62

To simulate the ACEI and angioedema dataset, we generated

(1) Categorical covariates $\boldsymbol{Z}$ that have the same mean and pairwise covariance as what are observed from the real data;

(2) Binary exposure X (ACEI = 1 and BB = 0) generated based on a logistic regression on the covariates (the propensity score model), using the coefficients observed from fitting the real data. For all cases, we hold the prevalence the same as the real data (65% ACEI). (**Table C 1**Error! Reference source not found. Observed Propensity)

(3) A pair of binary potential outcomes $(\mathrm{Y}(1),\mathrm{Y}(0))$ (angioedema within 30 days under exposure and control for the same subject) based on a logistic regression on the exposure and covariates (the outcome regression model), using the coefficients observed from fitting the real data. The observed outcome is thus $\mathrm{Y} = \mathrm{XY}(1) + (1 - \mathrm{X})\mathrm{Y}(0)$. For all cases, we hold the event rate in the control group (BB group) the same as the real data, which is equal to 0.05%. (**Table C 2** Observed Outcome Regression Model).

Table C 2. Relationship between outcomes given the exposure (ACEI and BB) and confounders

	Outcome Model (OR)	
	Observed Model	Observed+ Age*Diabetes Interaction
Confounders		
Heart Disease	1.13	1.13
Aspirin	1.38	1.38
Ischemic HD	1.07	1.07
OptRec Colon Surg	1.58	1.58
Allergic Reaction	1.54	1.54
Inpatient Hosp.	2.18	2.18
NSAIDS	0.93	0.93
Diabetes	0.73	--
Female	1.63	1.63
Age(Ref: 18-44)		
45-54	1.08	--
55-64	0.84	--
65-99	0.92	--
Age*Diabetes (Ref: 18-44 and Not Diabetic)		
45-54 & Not Diabetic		1.08
55-64 & Not Diabetic		0.84
65-99 & Not Diabetic		0.92
18-44 & Diabetic		0.73
45-54 & Diabetic		0.48
55-64 & Diabetic		0.37
65-99 & Diabetic		0.41
Exposure		
ACEI	2.51	2.51

In addition, we increase the strength of confounding by scaling up the coefficient in the propensity score model (multiply coefficients on the logOR scale by 1.5), while still holding the exposure prevalence and the baseline event rate the same (**Table C 1**). We also allow the propensity score model or the outcome regression model that generates the potential outcomes to include interaction terms between age and diabetes, to look at cases when the methods misspecify one of the models by missing the interaction (**Table C 1** and **Table C 2** Adding Interactions). Note that the outcome regression model on the propensity score hardly yields a function of covariates that matches the underlying data-generating model. So (mis)specification refers to methods that use regression models that actually fit on the covariates. For example, the regression on propensity score can use a misspecified propensity score model, which will make the propensity score be estimated with error and lose the balancing property to some extent; the augmented IPTW method can use misspecified propensity score model and/or misspecified outcome regression model.

We emphasize that the event rate is 0.05%, which yields about 50 angioedema incidences among the 100,000 subjects. We have twelve binary indicators including nine binary variables and three indicators for age category. When the exposure rate is 65%, for matching methods, there is going to be less control subjects than exposed subjects. We have also simulated the case when the exposure rare is 20%, which is more commonly seen, although it might result in a smaller matched dataset.

b. Results

Table C 3 shows the mean bias on the log(OR) scale, type I error, and power when the correct propensity score model and outcome regression model are specified. For matching methods, we calculate the bias assuming the ATT estimate.

In terms of bias, standardization with propensity score B-Splines performed similar to standardization with covariate adjustment. In settings when there is strong confounding effect with an exposure rate of 65%, or when the exposure rate is 20% under both moderate and strong confounding, standardization with propensity score B-Splines outperformed traditional methods, and also corrected the bias from adjusting for the propensity score as a linear term. However, when there is moderate confounding effect with an exposure rate of 65%, we observe similar biases between propensity score B-Splines and propensity score linear adjustment, and the augmented IPTW performs slightly better. The augmented IPTW had substantial increase of bias when the exposure rate is 20%, which is further from 50% compared to 65%. This is an evidence of sensitivity to inverting a propensity score that is closer to zero.

All methods had similar type I error and power except that matching had lower power. In particular, when the exposure rate is 20%, matching had much lower power, which can be due to a smaller size of the matched sample (at most 40% of full population matched). The valid type I error and high power showed that the direct estimation of variance is a fast and valid approach for inference.

Table C 3. Bias and Power in estimating the marginal OR by method ranging the strength of confounding and relationship between exposure and outcome

	Original Confounding				Strong Confounding			
	log(OR)=0		log(OR) = 0.92		log(OR)=0		log(OR) = 0.94	
Methods	**Bias***	**Type I error**	**Bias**	**Power**	**Bias**	**Type I error**	**Bias**	**Power**
Exposure rate = 65%								
Matching	-0.034	0.018	-0.023	0.894	0.012	0.023	0.027	0.896
Augmented IPTW	-0.010	0.017	-0.004	0.931	0.017	0.030	0.029	0.890
Standardization								
Covariates	-0.012	0.016	-0.010	0.938	0.015	0.024	0.019	0.884
PS Linearly	-0.013	0.018	0.007	0.942	0.014	0.029	0.044	0.888
PS Deciles	-0.030	0.016	-0.028	0.934	-0.005	0.024	-0.003	0.876
PS B-Splines	-0.013	0.016	-0.009	0.938	0.013	0.025	0.019	0.884
Exposure rate = 20%								
Matching	-0.060	0.030	-0.061	0.548	-0.023	0.030	-0.010	0.535
Augmented IPTW	-0.124	0.047	-0.046	0.740	-0.167	0.074	-0.069	0.674
Standardization								
Covariates	-0.082	0.044	-0.027	0.753	-0.075	0.064	-0.022	0.722
PS Linearly	-0.082	0.043	-0.037	0.743	-0.076	0.064	-0.049	0.701
PS Deciles	-0.095	0.043	-0.040	0.741	-0.094	0.065	-0.038	0.709
PS B-Splines	-0.083	0.043	-0.028	0.751	-0.078	0.069	-0.025	0.722

*mean bias on the log OR scale

Table C 4 shows the mean bias on the log(OR) scale, type I error, and power when the propensity score model is misspecified by missing the interaction term. **Table C 5** shows the mean bias on the log(OR) scale, type I error, and power when the outcome regression model is misspecified by missing the interaction term.

In both tables, we observed similar results in terms of bias as **Table C 3**, although misspecification of outcome regression model seemed to have more impact on the performance of the methods. The biases of standardization methods using either covariate adjustment or propensity score adjustment were slightly higher due to misspecification of models. The augmented IPTW method had less increase or even decrease in bias under model misspecifications, but the performance was unstable, due to large number of covariates and rareness of the outcome. Adjusting for propensity score deciles also had an unstable performance across all tables, having smaller bias under strong confounding and 65% exposure rate, but larger bias otherwise. The performance of adjusting for propensity score strata indicators or propensity score B-spines is sensitive to the degree of freedom determined by number of strata or number of B-spline basis functions. In practice, we suggest using cross-validation to select a valid number.

Type I error and power were similar comparing **Table C 4** and **Table C 3**. However, in **Table C 5** there was notable power loss under strong confounding. Again we see that matching had lower power than all other methods due to insufficient sample size.

Table C 4. Bias in estimating marginal OR when true propensity score model has interactions

	Original Confounding				Strong Confounding			
	log(OR)=0		log(OR) = 0.92		log(OR)=0		log(OR) = 0.94	
Methods	Bias*	Type I error	Bias	Power	Bias	Type I error	Bias	Power
Exposure rate = 65%								
Matching	-0.037	0.018	-0.026	0.900	0.017	0.026	0.030	0.891
Augmented IPTW	-0.014	0.018	-0.007	0.935	0.017	0.028	0.026	0.889
Standardization								
Covariates	-0.016	0.014	-0.012	0.946	0.016	0.026	0.018	0.881
PS Linearly	-0.017	0.016	0.004	0.948	0.016	0.031	0.042	0.885
PS Deciles	-0.031	0.015	-0.029	0.940	-0.003	0.024	-0.003	0.872
PS B-Splines	-0.016	0.015	-0.012	0.945	0.015	0.027	0.018	0.879
Exposure rate = 20%								
Matching	-0.062	0.029	-0.062	0.548	-0.029	0.029	-0.018	0.528
Augmented IPTW	-0.116	0.049	-0.042	0.745	-0.172	0.076	-0.069	0.668
Standardization								
Covariates	-0.083	0.043	-0.027	0.754	-0.077	0.067	-0.027	0.716
PS Linearly	-0.083	0.042	-0.037	0.746	-0.078	0.065	-0.054	0.691
PS Deciles	-0.095	0.041	-0.039	0.742	-0.096	0.066	-0.043	0.702
PS B-Splines	-0.084	0.043	-0.028	0.752	-0.080	0.070	-0.029	0.716

*mean bias on the log OR scale

Table C 5. Bias in estimating marginal OR when true outcome model has interactions

	Original Confounding				Strong Confounding			
	log(OR)=0		**log(OR) = 0.92**		**log(OR)=0**		**log(OR) = 0.94**	
Methods	Bias*	Type I error	Bias	Power	Bias	Type I error	Bias	Power
Exposure rate = 65%								
Matching	-0.047	0.017	-0.035	0.888	-0.156	0.026	-0.140	0.701
Augmented IPTW	-0.018	0.016	-0.011	0.925	-0.079	0.036	-0.070	0.747
Standardization								
Covariates	-0.020	0.015	-0.017	0.935	-0.081	0.033	-0.080	0.735
PS Linearly	-0.021	0.017	0.003	0.939	-0.087	0.041	-0.057	0.751
PS Deciles	-0.034	0.016	-0.031	0.932	-0.098	0.037	-0.097	0.720
PS B-Splines	-0.021	0.015	-0.016	0.935	-0.084	0.036	-0.080	0.736
Exposure rate = 20%								
Matching	-0.065	0.031	-0.065	0.543	-0.175	0.028	-0.154	0.340
Augmented IPTW	-0.126	0.047	-0.047	0.735	-0.264	0.105	-0.108	0.461
Standardization								
Covariates	-0.089	0.042	-0.030	0.746	-0.155	0.086	-0.054	0.531
PS Linearly	-0.090	0.042	-0.038	0.735	-0.152	0.086	-0.073	0.513
PS Deciles	-0.101	0.041	-0.041	0.733	-0.173	0.089	-0.069	0.517
PS B-Splines	-0.091	0.043	-0.031	0.744	-0.156	0.089	-0.056	0.537

*mean bias on the log OR scale

6. Discussion

In this paper, we have shown that there is great potential in using regression adjustment of the propensity score to estimate causal effects for rare binary outcomes, which fits in the postmarket drug surveillance research. We pointed out that although the propensity score is sufficient in balancing the confounders between exposure groups, regression adjustment directly using the propensity score as a covariate can result in bias, whereas a fast and simple correction of the bias comes from fitting flexible spline function of the propensity score.

Simulation study showed that flexible adjustment of propensity score in an outcome regression model resulted in less bias without loss of efficiency, and can outperform traditional methods when the propensity score model is correctly specified. When the propensity score was misspecified, regression adjustment of propensity score can have larger bias, but still performs comparably

to most methods. The augmented IPTW method performed better in such a situation, but might suffer from convergence problem with larger number of covariates due to rareness of the outcome. When the outcome regression model was misspecified, the augmented IPTW still had less bias, but sometimes very larger variance.

With non-rare exposure and a large cohort, if one is confident in doing a good job in fitting the propensity score model, we suggest fitting the propensity score model as a first step, and then use flexible regression adjustment of propensity score instead of using traditional propensity score methods such as matching and augmented IPTW. Although the augmented IPTW estimator can outperform all methods under strong confounding effect, in general, it requires fitting the outcome regression model and inversion of the propensity score, which makes it less stable. In addition, we also suggest fitting on propensity score deciles as a sensitivity analysis as it fits another nonlinear function of the propensity score and can perform well when the confounding effect is strong.