MINI-SENTINEL METHODS

IMPROVING SEQUENTIAL SAFETY SURVEILLANCE PLANNING METHODS FOR ROUTINE ASSESSMENTS THAT USE REGRESSION ADJUSTMENT OR WEIGHTING TO CONTROL CONFOUNDING

Prepared by: Jennifer C. Nelson, PhD,[1,2] Denise Boudreau, PhD, RPh,[1,2] Robert Wellman, MS,[1] Onchee Yu, MS,[1] Andrea J. Cook PhD,[1,2] Judith Maro, PhD,[3] Rita Ouellet-Hellstrom PhD, MPH,[4] James Floyd, MD, MS,[2] Susan R. Heckbert, MD, PhD,[2] Simone Pinheiro, ScD MSc,[4] Marsha Reichman, PhD,[4] Azadeh Shoaibi, PhD[4]

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

April 26, 2016

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel

Mini-Sentinel Methods

Improving surveillance planning for routine safety assessments that use regression adjustment or weighting to control confounding

Table of Contents

I. EXECUTIVE SUMMARY

The Food and Drug Administration's (FDA's) Sentinel Initiative aims to develop an electronic health care data system for proactive medical product safety assessments. To this end, Mini-Sentinel investigators have developed a suite of semi-automated tools to conduct routine multivariable risk estimation and sequential testing of an association between a pre-specified medical product and adverse event of interest. These tools use several different study designs and confounder adjustment strategies, including a self-controlled risk interval design, an exposure-matched cohort design with propensity score matching, or a full cohort approach with either regression adjustment for individual confounders or inverse probability weighting with a propensity score. Although these designs and statistical methods are well established in traditional epidemiological settings, their use for safety surveillance in a distributed health care database environment like Sentinel is relatively novel. More experience applying these methods in the Sentinel setting is needed to formulate best practices for conducting safety surveillance that ensures valid estimation. The purpose of this report is to provide suggestions about how to adapt the sequential aspects of surveillance planning to the Sentinel setting, using the routine tools that involve regression adjustment or weighting to control for confounding.

We first review current sequential surveillance planning methods, including prevailing guidance for randomized trials along with examples from the Vaccine Safety Datalink (VSD) project and Mini-Sentinel pilot, two national observational database safety monitoring programs. Based on this examination, we suggest several steps to improve future sequential surveillance planning in health care databases. These recommendations focus on sequential design selection, including sample size planning. Last, we illustrate these planning steps and present results of sequential analyses of two example associations: 1) angiotensin-converting enzyme (ACE) inhibitors and the risk of angioedema, and 2) angiotensin receptor blockers (ARBs) and angioedema.

This report has two main conclusions. First, existing methods used for sequential design planning for randomized trials and for observational safety surveillance assessments within the VSD and Mini-Sentinel provide a strong foundation upon which to build a more formal framework for future routine safety evaluations using electronic health care databases. Second, there are several ways that methods from randomized trials can be adapted to accommodate the unique challenges of conducting safety surveillance activities in the observational setting of electronic health record databases. There are also ways in which existing methods from observational settings like the VSD could be improved by further leveraging well established practices from trial settings and tailoring them to meet the challenges posed by an electronic data environment. Specifically, the working group proposed three simple planning steps to ease the development of the sequential design planning for future sequential surveillance in health care databases: i) use available data to inform a pre-specified sequential design and analysis plan in order to reduce assumptions and minimize later changes to initial plans (feasibility assessment); ii) describe existing uptake for the product of interest to determine whether or not there is adequate information to meet the (sample size) needs of the plan and conduct further, more resource-intensive planning steps, iii) statistically evaluate, jointly select, and clearly communicate the final sequential design and sample size considerations to all stakeholders in advance of implementation, and iv) implement the plan. To accommodate the dynamic and often unpredictable changes made to the database information by the health plans for administrative purposes, the working group also stressed the importance of preparing to be flexible in the implementation of initial plans and to document any resulting surveillance plan changes that may occur.

II. INTRODUCTION

A. STAGES OF POST-MARKET SAFETY SURVEILLANCE

Historically, post-market medical product safety monitoring has relied heavily on spontaneous reporting systems that contain passive and often voluntary reports from patients, health care providers, and other stakeholders who suspect that an observed adverse effect is related to a drug or vaccine that was given.[1-3] A big advantage of passive reporting is that results can be analyzed soon after release of the product to the market. Recognized disadvantages include the existence of strong reporting biases and an inability to conduct traditional statistical comparisons of risk (e.g., incidence rates between exposed and unexposed groups) due to lack of population denominator data. Other traditional sources of post-licensure safety evidence, such as a confirmatory randomized trial, can provide more reliable and accurate estimates of elevated risk. However, results from these resource-intensive assessments are not often available quickly after product licensure, involve restricted versus real-world populations, and may not provide adequate information about very rare events or adverse effects in certain subgroups. To overcome these limitations and complement existing surveillance tools, FDA undertook the Sentinel Initiative, which aims to create an active national surveillance system that leverages existing electronic health care data to proactively and rapidly assess medical product safety. [4,5]

The overall aims of post-licensure surveillance can be broadly classified into three main steps:[6,7]safety signal identification (or generation), refinement, and confirmation (or evaluation). Signal identification efforts are designed to detect new, previously unanticipated adverse events. Such evaluations often cast a wide net and assess hundreds or thousands of both specific and non-specific outcomes. Signal refinement investigations of a new product typically target, and may monitor over time, a small number of pre-specified potential adverse effects that are hypothesized to be of potential concern based either on previously generated signals from other data sources, biologic plausibility, or prior experience with products in the same class. A signal refinement surveillance activity typically does *not* involve specifying a detailed protocol with full adjustment for potential confounders tailored to each outcome. It instead implements a more basic surveillance plan that adjusts for key confounders to provide an improved estimate of risk compared to that from a signal generation exercise. Signal confirmation involves a more rigorous and in-depth assessment of a positive signal identified at an earlier stage that is intended to be more definitive through the development of a customized surveillance protocol tailored to a single product-outcome pair. In this report, we focus on signal refinement activities, which have been a major focus of the Mini-Sentinel pilot.

B. SIGNAL REFINEMENT IN SENTINEL

A variety of statistical methods have been developed for signal refinement activities within Sentinel. Specifically, four different approaches to risk estimation and testing of a potential association between a selected product and adverse event of interest have been developed in the framework of a semi-automated surveillance query tool, and these include adjustment for potential confounders.[8] These tools are complementary to one another as they use several different designs and confounder adjustment strategies:

1. A self-controlled risk interval design (SCRI) and within-person relative risk estimation
2. An exposure-matched cohort design, 1:1 or variable ratio propensity score (PS) matching, and conditional estimation of a odds ratio or hazard ratio (PS Matching)
3. A full exposed and unexposed cohort, logistic or Poisson regression adjustment for individual confounders, and estimation of an odds ratio or relative risk (Regression)

4. A full exposed and unexposed cohort, inverse probability weighted regression (with propensity score weighting), and unconditional estimation of a risk difference from a linear model (IPTW)

These tools can be used in a traditional one-time assessment (known in Sentinel as a Level 2 query) or they can be implemented repeatedly at multiple pre-specified time points, in a sequential monitoring framework (known in Sentinel as a Level 3 query). This latter setting allows prospective and routine monitoring of a new medical product over time as soon as there is adequate uptake of that product within the Sentinel population. Sequential risk estimation and testing methods are an appealing way to address concerns about potential drug-related adverse events because they allow data to be routinely evaluated as they are collected and raise a preliminary signal as soon as compelling evidence is observed. In this way, a decision to more deeply investigate a potential safety problem may be reached at a much earlier stage than would be possible with traditional analytical methods that conduct a single analysis after all the data have been observed. Sequential application, which is possible with any of these four tools, is thus very fitting for this purpose as it allows for repeated testing with Type 1 error rate control and potentially enables earlier identification of a possible safety signal as soon as sufficient information from new drug or vaccine recipients is available to detect elevated risks.

The designs and methods implemented by these monitoring tools are based on standard epidemiological and statistical approaches, and therefore, much is already known about their use in traditional observational study settings, including general advantages and limitations. However, the sequential application of these methods in a distributed database surveillance setting like Sentinel to monitor safety is relatively novel. Many practical and statistical challenges arise in Sentinel, and these necessitate methodological adaptations.[9] First, adverse events of interest to FDA are often uncommon, which means that standard methods based on large sample assumptions may not be appropriate. In addition, data are distributed across heterogeneous Data Partners, and individual level data cannot typically be pooled for analyses. Further, data are dynamic (especially for recently approved drugs), meaning that they are constantly being updated as new information is received into the Data Partners' administrative and clinical data systems. Last, there is a desire for early detection, which motivates the use of sequential monitoring to identify potential safety issues as soon as sufficient evidence is available. To complicate matters further, in a sequential surveillance setting, early adopters of new medical products may differ from those who eventually use the product in ways that cannot be anticipated in advance. Although Mini-Sentinel investigators have successfully modified existing sequential methods to overcome some of these challenges,[10-12] more experience applying these methods in the Sentinel setting is needed to formulate best practices and ensure valid estimation. In this report, we focus on improving practices for surveillance planning for the Regression and IPTW query tools. Many of the outlined surveillance practices are also applicable to other methods.

C. PURPOSE AND SCOPE OF THIS REPORT

Once the FDA has selected a particular product-outcome pair of interest and determined that Sentinel is an appropriate environment to conduct a safety assessment, it is important to understand what steps are needed to develop a detailed surveillance plan using the available tools. The purpose of this report is to provide suggestions about how to plan a safety assessment using the two tools that estimate exposure-outcome associations adjusted for confounders using either regression or weighting in the Sentinel setting. These tools can be applied as a one-time (Level 2) analysis or conducted repeatedly at multiple pre-specified points in time, using a sequential monitoring framework (Level 3). A plan involving one analysis may make sense to assess safety for an existing medical product that has been on the market for many years, has had considerable use in the Sentinel population, and so is already well-

powered to address the question of interest. This was the case for the two examples provided in Section V. A plan involving multiple assessments of the data over time could be employed for a newly marketed product or an existing product approved for new indications, with analyses conducted routinely as adequate amounts of new product uptake occurs so that safety signals of concern could be detected earlier.

The scope of the planning recommendations in this report assumes the following:

- FDA has already determined that the safety question of interest (i.e., the product-outcome pair) is suited for examination using Sentinel data
- The study design taxonomy has been consulted to help suggest an appropriate epidemiological study design or designs (i.e., self-controlled or a cohort design), and
- Either the Regression or IPTW regression tool has been preliminarily deemed appropriate.

In other words, this report does *not* address earlier stage planning questions that are also important:

1. What safety questions should be considered and prioritized for safety assessments using Sentinel data versus other systems?
2. How do we decide which design (or query tool) is most appropriate for a given product-outcome pair?
3. Among cohort designs, how do we decide which confounder adjustment method is preferred?

Answering question 1 above about what safety questions should be prioritized relates to the fitness-for-purpose of Sentinel data.. In general, Sentinel will be a reasonable environment to use if the outcome and product of interest are captured adequately by Sentinel Data Partners, the occurrence of outcomes and use of products can be accurately defined using data elements available in the Sentinel Common Data Model (CDM), and the key sources of potential confounding can be measured well using variables available in the CDM. Suggestions on question 2, the question about choosing an appropriate surveillance design for a selected product-outcome pair, have been developed by a previous Mini-Sentinel workgroup and can be found elsewhere. [13,14] To address question 3, the question on confounder adjustment method choice for cohort designs, the advantages and limitations of each approach should be weighed. Although we do not provide a detailed discussion on this topic here, **Table 1** highlights some of these considerations. Note that these features are specific to the three cohort tools that have been developed for use in Sentinel.

Table 1. Features of Sentinel's available confounder adjustment methods for cohort designs

Method	Advantages	Limitations
Propensity Score (PS) matching	• Estimates a hazard ratio (HR), a familiar quantity that is well understood by epidemiologists • Can have a simple and intuitive interpretation • Is applicable for chronic medication use or longer-term outcome follow-up (e.g., occurring within months versus within days or weeks of drug initiation), which is the relevant setting for most drugs • Can adjust for a large number of	• Loses available adverse events/ information/power by sampling just a subset of available comparators • Loses available adverse events/ information/power for a patient when its matched pair is censored* • Requires upfront effort/decisions (to conduct matching) to restrict to a matched subset of the full cohort • Is complex to implement repeatedly over time due to dynamic changes in

Method	Advantages	Limitations
	confounders, even with rare events	data (e.g., re-matching issues) • Requires an adequate number of exposed and unexposed patients (for PS estimation) • Does not account for differences in PS variability across Data Partners • Less easy to conduct subgroup analyses • Need to trim or restrict to avoid including patients very unlikely or likely to receive exposure
Regression (with separate confounders or using a summary score like a propensity score)	• Uses all adverse event information from available cohort members (i.e., from exposed and comparators) • Estimates a relative risk (RR) or odds ratio (OR), quantities that are well understood by epidemiologists • Is a flexible adjustment method (i.e., can use a PS or individual variables and using different functional forms) • Is applicable for chronic medication use and/or longer-term outcome follow-up (e.g., >1 month), often relevant scenarios for drugs • Can adjust for a large number of confounders if PS is used to adjust • Easy to conduct subgroup analyses	• Requires an adequate number of patients with and without an adverse event • If adjustment for individual confounders is desired, -Number of confounders is limited if events are rare -Number of confounders is limited by the need to ensure a de-identified aggregated dataset (see Section IVD) • Need to remove/exclude outlying observations
Propensity score weighting	• Uses all adverse event information from eligible cohort members (i.e., from exposed and comparators) • Estimates a risk difference (primary), a quantity that is often of interest to policy decision-makers. Relative risks can be approximated but not what method is designed to estimate. • Can adjust for a large number of confounders, even with rare events • Directly accounts for differences PS variability across Data Partners • Is more stable and thus generally more powerful than methods that make relative comparisons of risk	• Requires an adequate number of exposed and unexposed patients (for PS estimation) • Less well known compared to other confounder adjustment methods • Is designed for short-term exposure and acutely occurring outcomes (i.e., not applicable for chronic drugs and long-term follow-up) • Need to trim or restrict to avoid including patients very unlikely or likely to receive exposure (i.e., those with large weights that can unduly inflate the variance)

Method	Advantages	Limitations
	• Can permit a causal inference interpretation, as in randomized trial[15] • Easy to conduct subgroup analyses	

*Applies to 1:1 matching and to a lesser extent variable ratio matching when conditioning on the matched pair or set

In Section III of this report, we review existing practices that have been used to plan for similar sequential monitoring activities in other settings. These include the VSD, ongoing Mini-Sentinel assessments, and sequentially monitored randomized trials. We describe important feasibility questions and the type of feasibility data that should be collected in advance of surveillance plan development to make planning more efficient and informed. Our primary emphasis, however, is on developing a suggested framework for sequential design selection for the Regression and IPTW methods, which we cover in Section IV. This includes practical suggestions on using available data to inform sample size planning. Although the focus of this report is on planning steps needed when using the Regression and IPTW methods, many of the suggestions we provide are also applicable and can be readily extended to the SCRI and PS Matching tools. In Section V, we apply these planning steps when possible and report results from a sequential analysis using the Regression and IPTW methods using the associations between ACE inhibitors and angioedema as well as ARBs and angioedema as related examples. Note that we were limited in our ability to apply the many of the planning steps to the example analysis in Section V because the planning steps were developed by this work group at the same time that the data were extracted and plans were made for the sequential analysis in Section V.

III. REVIEW OF EXISTING SEQUENTIAL DESIGN PLANNING PRACTICES FOR SAFETY SURVEILLANCE

In this section, we summarize approaches for planning sequential assessments that have been used in prior surveillance activities or studies. We focus specifically on the sequential aspects of planning and do not cover more traditional epidemiological and clinical planning decisions (e.g., defining the population, selecting outcomes, choosing comparators, etc.) that need to be made since these steps are already described in the available PROMPT Users' Guide. [8] We first highlight best practices from sequential randomized trial settings, where sequential methods have been used extensively,[16] as some of these practices may also be useful to consider for observational safety assessments. We then describe the experience of the VSD, which has since about 2005 conducted sequential vaccine safety surveillance in a distributed observational database setting similar to Mini-Sentinel.[17] Last, we describe key sequential planning steps undertaken in several pilot surveillance activities conducted within Mini-Sentinel. We begin with a brief overview of sequential methods and a description of the type of planning activities that are often undertaken.

A. OVERVIEW OF SECUENTIAL DESIGN METHODS

Sequential methods are designed to repeatedly test a hypothesis based on data as it accumulates over time. The overall false positive (Type 1) error rate across the multiple tests is pre-specified and controlled. Early rejection of the null hypothesis is possible based on preset decision rules. Sequential testing has been extensively used for many decades in randomized trials. More recently these methods have been adapted for use in observational safety settings, including in surveillance activities using electronic health care data. An example of this approach being used in an observational safety context is the Mini-Sentinel activity to assess whether acute myocardial infarction risk is elevated among users of a

new oral antidiabetic drug, saxagliptin, as this new drug is taken up, compared with other medications.[18] Sequential monitoring is attractive for post-market drug and vaccine safety surveillance because it can lead to detection of a potential safety concern as soon as pre-specified criteria for an elevated adverse event risk are met. Several continuous sequential methods have been proposed for use in observational safety surveillance, including cumulative sum charts,[19,20] maximized sequential probability ratio tests (MaxSPRT),[10] and sequential generalized likelihood ratio (GLR) tests.[21] Group sequential methods, which are widely used in randomized trials and involve less frequent interim testing, have also been adapted for use in observational safety settings like Mini-Sentinel. [11,12,22-25] Standard group sequential methods used in randomized trials, such as error-spending,[26] may also be usable in safety surveillance in situations where adverse events are more common and large sample statistical assumptions are met.

Planning a sequential observational safety surveillance assessment involves both the standard activities that are typically undertaken for an epidemiological study with a single analysis at the study's end (e.g., defining the eligible population, choosing appropriate comparators, selecting a suitable outcome risk window, identifying potential confounders, etc.) as well as additional considerations related to the use of a sequential design:

1. When should surveillance start and end?
2. How frequently should interim tests be performed?
3. What should the threshold be for a safety signal and should it change over time?

Answers to these questions determine the design's statistical properties (e.g., Type 1 error, sample size and power, and expected time until signal detection). Frameworks to address these questions in randomized trial settings are well-established, and decisions are typically guided by the trial's scientific goals, ethical concerns, and practical circumstances. [27] Less consideration has been given to which sequential designs may be preferred and how they should be selected in an observational safety setting, where the scientific safety questions, consequences of a signal, and costs of false positive and negative errors differ. Establishing a systematic process for planning sequential surveillance in an observational database setting is important as it could yield important improvements over existing surveillance planning practices, such as increased power to detect serious rare adverse events.

B. BEST PRACTICES FROM RANDOMIZED CLINICAL TRIALS

FDA provides extensive guidance on statistical principles for clinical trials conducted by industry, many of which involve group sequential interim monitoring.[28] In addition, a set of minimum standards for adaptive randomized clinical trials has been recently developed for comparative effectiveness research conducted within the Patient Centered Outcome Research Institute (PCORI).[29] Some of these principles may also be applicable to sequential studies that are not adaptive and are worth considering. **Table 2** below summarizes key recommendations from both these sources. We discuss the relevance and applicability of these recommendations for observational safety settings like Sentinel in Section IV.

Table 2. Key recommendations from FDA and PCORI on the conduct of sequential trials

Recommendation	Description
Pre-specify statistical design and primary analysis and document changes	All statistical methods should be pre-specified prior to obtaining information on treatment outcomes, including the schedule of interim analyses, stopping rules and their properties, primary hypotheses, underlying statistical model, use of 1- versus 2-sided tests, and designation of primary versus exploratory analyses. It is important to document protocol deviations as changes made to the original plans can weaken and even invalidate the results.

Recommendation	Description
Evaluate statistical properties of the design in advance	The statistical properties of the design should be evaluated *a priori* so that they are understood prior to implementation and in the context of the research question (e.g., adequate power for several assumed true treatment effects). For complex designs, this might include evaluating properties over a range of assumptions relating to size of treatment effect, missing data, dropout rates, etc. Technical details be included in an appendix (e.g., statistical models and thresholds for the primary analyses along with calculation details or software used, operating characteristics for the design along with methods and assumptions for computing them (e.g., if based on simulation).
Communicate and vet the design in advance	The sequential design and analyses should be clearly communicated and vetted with key stakeholders to assess acceptability to address the primary aims.
Account for multiple testing	The chance of making a Type 1 error may increase due to testing multiple outcomes, treatment comparisons, subgroups, or repeated analyses over time and should be addressed, potentially using frequentist Type 1 error adjustment methods.
Interpret exploratory analyses with caution	Exploratory analyses (e.g., subgroups) should be interpreted with caution and should generally not be used to make definitive conclusions regarding treatment effects.
Ensure proper oversight and reporting	Proper statistical oversight of trial conduct should be in place, and reporting of the results should be done in a consistent fashion.

C. EXPERIENCE FROM THE VACCINE SAFETY DATALINK COLLABORATION

1. Continuous sequential monitoring

After preliminary exploration with the original sequential probability ratio test (SPRT),[30] early sequential safety surveillance within the vaccine safety datalink (VSD) primarily utilized the MaxSPRT method. This approach involves near-continuous sequential monitoring. It uses a 1-sided likelihood ratio test (LRT) that rejects the null hypothesis of no difference in the risk of a pre-specified adverse event between a vaccine of interest and comparator if the log likelihood ratio (LLR) exceeds a constant upper value. In other words, MaxSPRT uses a constant (or flat) signaling boundary over time on the scale of the LLR. Surveillance using MaxSPRT was typically conducted for a small number of pre-specified outcomes (about 5-10) for about two or three years following introduction of a new vaccine [31-36] or, in the case of influenza vaccine monitoring, for the duration of influenza season.[37,38] In some instances, statistical power was computed *post hoc* after surveillance was completed. [39]

Continuous testing is advantageous because, on average, it can detect true safety signals sooner.[40] However, continuous testing is inherently less powerful than if testing is less frequent, given a fixed sample size. This is because more frequent testing increases the overall chances of producing a false signal or Type 1 error. To maintain the same Type 1 error, the signaling threshold must be increased when testing is more frequent, which in turn reduces power. In addition, continuous testing is not currently feasible within Sentinel since new data updates by Data Partners are not conducted in real-time, continuous fashion. In particular, the signaling threshold for a continuous testing procedure assumes that the hypothesis test of interest is conducted as each new observation occurs in the population. In Mini-Sentinel, new data updates at each Data Partner are typically conducted on a quarterly basis. Thus, newly updated data are available in batches (i.e., not as each new event occurs),

and so testing is only feasible on an interim basis as each new batch of data is available. One could conduct interim testing as each new batch accrues and apply the continuous testing threshold, but this signaling threshold would be unnecessarily conservative, yielding suboptimal power and smaller than desired Type 1 error.

A flat boundary generally yields a lower signal threshold at early testing time point, which may also enhance early detection of true signals. However, by *not* employing early conservatism, use of a flat boundary can also generate false positive signals based on relatively little information at early analyses due to small sample variability. This problem was observed in several initial VSD studies [41] and led to the development of continuous methods that implement a 'delayed start', which involves postponing the first test until a certain number of events are observed.[42] Additional properties of continuous compared to group sequential testing methods in a post-licensure safety setting have been described and evaluated previously. [10,40,43,44]

2. Group sequential monitoring

Group sequential methods were first adapted from clinical trials for use in an observational safety setting in a VSD study of a new pentavalent combination vaccine for infants (trade name: Pentacel). [22] Tseng et. al. also used a group sequential approach to monitor 13-valent pneumococcal conjugate vaccine (PCV13) safety in children.[45] Similar to prior VSD studies, the Pentacel safety study used a 1-sided LRT with a flat signaling boundary to test whether the risk of several targeted adverse events was elevated among Pentacel recipients versus comparators. Instead of continuous testing, however, 12 group sequential interim tests were planned. The first test, which occurred after 1 year of Pentacel uptake (N=33,308 doses), was purposely delayed to apply early conservatism and minimize early false positive signaling. Subsequent tests were planned to be equally-spaced based on the number of newly accumulating doses of Pentacel needed to achieve specific statistical power goals (i.e., spacing between analyses was based on the amount of available information or 'information time' as opposed to spacing that is based on a preset number of weeks or months in 'calendar time').

Given this sequential design (i.e., given this schedule for testing and the flat boundary choice) and the expected adverse event rate among comparators, the maximum total sample size required to achieve at least 80% power to detect a specific minimum relative risk of interest for each outcome was computed. For more common events, this resulted in tests (subsequent to the first) being performed after each additional batch of 3,500 doses of Pentacel was observed, up to a maximum sample size of about 72,000 doses. For less common events, tests (subsequent to the first) were planned after each new 10,500 doses accrued among VSD enrollees, with a maximum sample size of about 150,000 doses. For the most rare adverse events (i.e., <0.05 cases per 10,000 doses), event counts were tracked but no formal sequential monitoring was planned. In addition to monitoring pre-specified adverse events, a non-specific severe outcome (any-cause hospitalization) and several control outcomes were analyzed as end-of-study (not sequential) endpoints.

In settings like the VSD and Sentinel, the data are not only observational (versus a controlled clinical experiment), but they are collected for reasons other than surveillance or research. They are captured and dynamically updated over time by health care organizations and health plans for administrative and clinical purposes. As a result, many unanticipated occurrences and changes in the data occur during the surveillance period. These can both impact the expected variability in the adverse events over time as well as constrain our ability to conduct sequential analyses exactly according to our pre-specified plan. Here are examples from the Pentacel study of complications that can arise:

1. There was unanticipated differential uptake of the Pentacel vaccine by age and by Data Partner.
2. Each planned interim analysis could not be performed at exactly the number of doses that was pre-specified because data were not continuously refreshed but rather updated in groups of newly added doses at discrete (weekly) intervals over time. For instance, the second analysis was planned to occur at 36,808 doses. However, it was actually conducted at 37,851 doses in week 59 of surveillance because fewer than the required 36,808 total doses were available at week 58 and more than 36,808 accrued by the following week.
3. Due to an unforeseen data quality issue that was identified and later corrected, an unexpectedly large amount of previously missing data accrued at a single time point from one Data Partner.

This lack of experimental control and unexpected data occurrences and changes over time impact the adverse event variability and, in turn, the probability of committing a Type 1 error that investigators want to control. To account for these unpredictable features or changes in the data in the Pentacel analysis, the sequential boundaries were updated modestly over time. Specifically, to maintain proper error control, the planned boundaries were adjusted at each analysis to account for the actual (versus planned) way in which the data accrued. The actual departures from the planned analysis time points were not sizable, however, and the boundary adjustments were thus correspondingly small.

As in the Pentacel study, actual conduct of the PCV13 safety study was modestly different than initially planned. In particular, investigators planned to finish surveillance for all pre-specified outcomes within two years, before the end of the VSD contract period. However, accrual of information for the rarest events did not occur quickly enough to meet this deadline. Thus, some testing plans were modified in the end so the study could be completed in the required time frame. In addition, when a signal was detected for Kawasaki disease at the second group sequential test, investigators continued to descriptively monitor the additional new cases of this disease as formal testing for other outcome of interest continued for PCV13.

A summary of the main features of the continuous and group sequential designs that were developed and implemented in the VSD is in **Table 3**.

Table 3. Summary of the planned continuous and group sequential designs used in the VSD

Sequential design features	Continuous testing using the MaxSPRT approach	Group sequential LRT in the Pentacel safety study	Group sequential LRT in the PCV13 safety study
Surveillance start	Conducted as soon as uptake begins (i.e., after week 1)	Delayed start, after 1 year of uptake to employ early conservatism	Specified in information time (number of doses) and based on power to detect specific RRs
Surveillance end	Specified in calendar time ~2-3 years after the first dose	Specified in information time based on power to detect specific RRs; depends on event prevalence (N=72,000 doses for common and 150,000 for rare events)	Specified in information time based on power to detect specific RRs; depends on adverse event prevalence
Frequency of testing	Specified in calendar time as weekly	12 total tests based on information time; spacing depends on event prevalence (N=3,500 or 10,500 doses between each analysis)	12 total tests based on information time; spacing depends on event prevalence
Duration of	Specified in calendar	Specified in information time;	Specified in information time;

Sequential design features	Continuous testing using the MaxSPRT approach	Group sequential LRT in the Pentacel safety study	Group sequential LRT in the PCV13 safety study
surveillance	time as 2-3 years	resulted in ~2.5 years	resulted in ~2 years
Shape of signaling threshold over time	Constant (flat) threshold on the scale of the LRT scale	Constant (flat) threshold on the scale of the LRT test statistic	O'Brien-Fleming threshold on the LRT scale; higher at earlier analyses to be conservative
Test statistic	LRT	LRT	LRT
Sidedness of test	1-sided	1-sided	1-sided
Thresholds adjusted over time?	No	Yes	No
Apply data lag?	2-3 months	2-3 months	2-3 months
Freeze prior data?	Freeze results from prior analyses and only add new information	Primary: Cumulatively refresh all data since start of surveillance at each new interim analysis Secondary: Freeze results from prior analyses and only add new data	Cumulatively refresh all data since start of surveillance at each new interim analysis

The final two rows of **Table 3** address two technical data-related questions that sequential surveillance plans face in the VSD. First, should data be lagged? In other words, instead of including all data that have been captured in the health care databases up to the day before an interim analysis is conducted, should we wait several weeks or months before including data on a given patient in an analysis to increase the probability that all relevant information (i.e., on exposure, adverse events, and confounders) has been correctly and completely captured in the database? Second, at each interim analysis when we examine cumulative data since the beginning of the surveillance period, should we freeze the previously analyzed data from the prior analyses and only add new data that has been captured since the prior analysis? Or, should we cumulatively refresh all the information we have observed since the beginning of the study?

With regard to the first question, the standard protocol for sequential safety studies within the VSD has been to lag the incoming data for analysis by about 2-3 months. (Note: One exception is influenza vaccine surveillance, where accessing data in real-time without a lag is essential due to the short duration of influenza season. Special methods have been proposed to accommodate the partially-accrued data in this situation. [39] For instance, if an analysis is conducted on March 1, the most recent data included in that analysis would be those observed through January 1. This lag period was instituted because some relevant vaccine and adverse event information is known not to be captured in the health plan databases instantaneously (e.g., relatively slower-arriving claims data when enrollees are seen at hospitals outside the integrated health system Data Partner). The rationale for waiting 2-3 months is that VSD data have been documented to stabilize dramatically and become much more complete after this time period. Having relatively stable data is important when conducting sequential tests since each new test conditions on the prior information. The approach to freezing (or not) prior data has varied by

VSD study, depending on specific design and method considerations. In some cases, multiple approaches were used to assess the impact of these different strategies on the final results.

D. PRIOR AND ONGOING SAFETY ASSESSMENTS IN MINI-SENTINEL

A small number of sequential safety evaluations for drugs [18,46] as well as vaccines[47] have been conducted or are currently ongoing within Mini-Sentinel. Many of the lessons learned from sequential safety studies conducted within the VSD were applied when planning these surveillance activities. A summary of the key features of several designs that have been developed and implemented in Mini-Sentinel are in **Table 4**. All involved group (versus continuous) sequential designs since new data updates by Data Partners are not conducted in real-time, continuous fashion within Sentinel.

Table 4. Summary of the main traits of sequential plans used in selected Mini-Sentinel activities

Sequential design features	Saxagliptin	Rivaroxaban	Influenza vaccine
Surveillance start	Specified in information time and based on power to detect specific HRs; resulted in 1st analysis at 2 years after July 2009 licensure	Specified in information time to occur at 35% of the total person-time and based on power to detect specific HRs	Conducted as soon as a pre-specified minimum number of events occurred following projected start of influenza vaccine distribution
Surveillance end	Specified in information time and based on power to detect specific HRs; resulted in last analysis ~6 years after licensure	Specified in information time and based on power to detect specific HRs; surveillance is ongoing	Specified to coincide with end of influenza season; based on expected # of events (using historical data) during influenza season, inflated slightly to prevent ending prior to the end of the season
Frequency of testing	7 total tests, planned to be equally-spaced based on information time	5 total tests, planned based on information time to occur at 35, 47, 62, 80, and 100% of the total person-time and designed to coincide with expected quarterly tests	Once minimum event threshold is reached, refreshes by Data Partner were roughly quarterly but staggered so new data appeared monthly.
Duration of surveillance	Specified in information time; resulted in ~6 years	Specified in information time; surveillance is ongoing	Specified to coincide with the duration of influenza season
Shape of signaling threshold over time	Constant (flat) threshold on the scale of the Wald test statistic	Constant (flat) threshold on the scale of the Wald test statistic	Constant (flat) threshold on the scale of the LRT test statistic
Test statistic	Wald	Wald	LRT
Sidedness of test	1-sided	2-sided	1-sided
Thresholds adjusted over time?	No	No	No
Apply data lag?	Varied by Data Partner (some lag data by 6-9 months, others do not	Varied by Data Partner (some lag data by 6-9	Used historical information on data lags for each Data Partner to determine when

Sequential design features	Saxagliptin	Rivaroxaban	Influenza vaccine
	implement any lag)	months, others do not implement any lag)	data were expected to be ≥85% complete.
Freeze prior data?	Cumulatively refresh all data since start of surveillance but preserve matches from prior analyses whenever feasible	Cumulatively refreshed data since start of surveillance. Explored different methods for matching, ranging from retaining prior matches to re-matching all data at each analysis	Freeze prior analytic results but cumulatively refresh all data since start of surveillance and incorporate any new data by appending it to the prior analytic dataset

As described for the VSD studies in the previous section, the actual sequential conduct of pilot Mini-Sentinel evaluations was not always the same as specified in initial plans. For instance, in the rivaroxaban surveillance activity, statistical power was estimated for various potential scenarios of interest (e.g., varying minimum detectable hazard ratios (HRs) of interest) assuming that five group sequential analyses would be conducted based on information time when 35, 47, 62, 80, and 100% of new users needed based on sample size calculations were observed, respectively. Thus, planning for interim tests was done on an information time scale. Based on these calculations, the maximum sample size (i.e,. the sample size at the fifth and final planned analysis if no safety signal is detected) required to achieve 80% power to detect a HR of 1.5 for the least common outcome (of intracranial hemorrhage) was estimated to be about 16,000 new rivaroxaban users. In practice, however, sequential testing was conducted at convenient time points primarily based on practical considerations. Specifically, the first test was conducted as soon as possible in calendar time after the surveillance plan was finalized, and subsequent tests were planned to occur quarterly, after all Data Partners had an opportunity to refresh their databases. This resulted in a first analysis that included about 15,000 new rivaroxaban users, which was very close to the desired maximum sample size after all five analyses. In this case, the planned sequential design (based on information time spacing between interim analyses) and the actual implementation of these analyses (based on practical considerations) did not coincide as initially hoped.

E. SUMMARY

This prior work points to several important aspects of surveillance activities that use a sequential design within Sentinel that should be addressed during the planning phase for sequential safety surveillance activities:

- Pre-specification of surveillance design and analysis plans:
  - Primary versus secondary (or sensitivity) analyses
  - Sequential versus one-time analysis
  - Multiple testing
  - Subgroup analyses
- Use of existing data (Step 1) to inform surveillance planning and reduce the number of assumptions that need to be made at the planning phase (e.g., related to sample size estimation, planning the timing of sequential analyses, etc.)
- Clear communication (Step 3) of the sequential design and analysis properties to the surveillance team to facilitate transparent design selection and evaluation, to include:
  - Desired duration of surveillance (in calendar time)

    - Desired sample size for surveillance (in information time – i.e., expected events)
    - Interim testing plan (number and timing of analyses)
    - Signaling threshold level over time
- Preparing to be flexible and to document any resulting changes to initial plans caused by:
    - Unpredictable uptake rate and population composition of new exposure
    - Incomplete data that are dynamically updated over time
    - Alignment of information time and calendar time preferences
- Interpretation and consistent reporting of results

IV. SUGGESTIONS ON SEQUENTIAL SURVEILLANCE PLANNING WITHIN SENTINEL

A. OVERVIEW AND GUIDING PRINCIPLES

In Section III, we reviewed methods for planning sequential evaluations that have been used previously in randomized trials and observational safety surveillance, focusing on the sequential (as opposed to the epidemiological) aspects of decision-making. In this section, we provide suggestions for planning future sequential evaluations using the Regression and IPTW modules that build on this prior background. Recall that these planning recommendations assume the following:

1. FDA has already determined that the safety question of interest (i.e., the product-outcome pair) is suited for examination using Sentinel data
2. The taxonomy report has been consulted to help suggest an appropriate epidemiological study design or designs (i.e., self-controlled or a cohort design), and
3. Either the Regression or IPTW regression tool has been preliminarily deemed appropriate.

There are three important additional features of our proposed suggestions. First, we focus on sequential design recommendations, which is step 4 in the overall User Guides' documentation:[8]

1. Select and define health outcomes of interest
2. Define exposure and cohort eligibility
3. Select comparator
4. Specify sequential surveillance plan
5. Implement sequential surveillance (report results, modify plan as needed)
6. Follow-up on signals

However, we also comment on how these sequential planning steps fit in a broader planning context for either a one-time analysis or a routine monitoring assessment (**Table 6**). Second, to make this work practical and concrete, we present these suggestions in the form of a checklist of planning steps with rationale for each step. More detail is provided for planning steps that are more involved, specifically confounder selection (Section IVD) and sequential design selection (Section IVE). Last, our recommendations are driven by several core guiding principles for surveillance in general, which are based on Sentinel's central values. In particular, we desire surveillance planning that is:

- Simple -- so planning can be rapid, efficient, and scalable
- Interpretable -- so planning steps are easy to understand and repeat
- Transparent – so planning decisions can be easily shared with relevant stakeholders
- Scientifically sound – to ensure rigorous surveillance that leads to maximal public health benefit

B. INCORPORATING LESSONS FROM PRIOR SEQUENTIAL EVALUATIONS

We determined that many of the established planning practices for randomized trials are also important to maximize the integrity of a sequential safety evaluation in an observational surveillance setting like Sentinel. However, the extent to which each recommendation applies may vary due to practical and scientific differences between the clinical trial setting and population (**Table 5**).

Table 5. Relevance to Sentinel of FDA and PCORI recommendations on the conduct of sequential trials

(i.e., relevance to observational post-marketing safety surveillance settings)

Recommendation from randomized trials	Relevant for observational surveillance?
Pre-specify statistical design and primary analysis and document changes	Yes. It is equally important in observational settings to pre-specify analytic plans to the extent possible. However, observational surveillance is subject to many more unknowns and may need to flexibly accommodate some changes when plans cannot be implemented as initially expected. Such changes should be documented and explained so that appropriate interpretations may be made.
Evaluate statistical properties of the design in advance	Yes. But it may not be as desirable or practical to conduct an extensive performance evaluation for surveillance applications because: i) Surveillance may be done for *many* exposure-outcome pairs at once, making it less feasible to conduct an extensive evaluation for each design, ii) Many unknowns can lead to changes in the actual versus designed implementation, which may down-weight the need to understand the planned design's performance in depth. It also may be helpful to use relatively simple designs that are well understood, can be re-used, and can be scaled up.
Communicate and vet the design in advance	Yes. It is vitally important that key stakeholders, especially FDA, understand how the design will work in practice so any actions taken based on a generated safety signal (when used in combination with all other available safety information) are suitable.
Account for multiple testing	Yes. However, the importance of strict accounting for random variation via multiple testing may be less in an observational surveillance setting since systematic variation will be (relatively) larger and sample sizes relatively larger. It is likely worth adjusting for sequential tests across multiple analysis time points but it may be less necessary to adjust across multiple outcomes (since very few outcome are targeted for surveillance) or subgroups (since this is already designated as exploratory)
Interpret exploratory analyses with caution	Yes. In general, surveillance results are more exploratory than results from trials. However, when pre-specified, primary surveillance results may reasonably test specific hypotheses. Results of surveillance analyses that are *not* pre-specified should be considered as hypotheses for further evaluation.
Oversight and reporting	Yes. Statistical oversight and reliable reporting are key components for surveillance, given the data and analysis complexities and the desire for transparent presentation.

The sequential vaccine safety surveillance experience within the VSD and the pilot surveillance activities conducted within Mini-Sentinel offer further lessons that should be considered when planning future safety surveillance activities within Sentinel:

- **Preliminary data:** Assessing the rate of initial uptake of the new drug or vaccine prior to developing the surveillance plan is extremely valuable for many reasons:

  - estimating the sample size needed to adequately address the safety question
  - determining the expected rate of uptake and how quickly sample size requirements may be achieved in calendar time
  - identifying which Data Partners are experiencing the most new uptake and can thus provide information relevant to the safety question of interest

- **Preliminary discussion**: Clear communication and joint selection with FDA of a sequential design's operating characteristics in advance is essential so that the meaning of a safety signal is well understood if it should occur.
- **Surveillance start:** Using a sequential boundary that employs some early conservatism (e.g., a delayed start or higher boundary at earlier versus later analyses) can help reduce the generation of false positive signals based on relatively little information at early analyses due to small sample variability.
- **Surveillance end:** Conducting a traditional sample size calculation is helpful to understand how much data is needed (which, in turn, determines how long it is necessary to conduct surveillance) to address a particular safety question of interest.
- **Timing of analyses**: Implementing interim analyses in an unpredictable observational setting introduces a need to be flexible. It often makes sense to plan the timing of interim analyses based on information (e.g., expected number of events) to estimate power and to ensure that there is adequate new data at each analytical time point to warrant carrying out the surveillance plan. However, since the rate of new drug uptake is not known and since Data Partner data updates are performed periodically in calendar time, we need to be prepared to adjust initial plans based on actual uptake and calendar time constraints.
- **Dynamically changing data**: Implementing a time lag between when data are first captured by a Data Partner and when they are included in an analysis is important to ensure more complete capture of information at any given analytical time point and to increase stability in the dataset across analyses. This can be implemented in Sentinel using the existing Cohort identification and Descriptive Analysis (CIDA) tool when data are pulled for analyses. Ideally, the size of the lag would be set to ensure that most (e.g., 90%) of the data needed for the analyses would be expected to be complete. This could be 3-9 months, depending on the Data Partner.

C. PROPOSED PLANNING FRAMEWORK: 4 STEPS, TIMELINE, AND RESULTS

Based on these guiding principles and lessons learned from prior work, in this section we propose an overall surveillance planning framework for Sentinel. This suggested framework can be used both for safety evaluations involving a single (Level 2) or sequential (Level 3) analyses, for established products that have been on the market for many years or for newly marketed ones, and for drugs as well as vaccines. Specifically, we propose that these following basic steps be implemented as soon as a product is identified as a priority for safety assessment within Sentinel:

Step 1: Conduct a descriptive quantitative *Feasibility Assessment* designed to facilitate the development of a preliminary surveillance plan. This step does *not* involve assessing product-outcome relationships. After completing this step, the surveillance team will be able to:

- Provide draft entries for the required fields on the Level 2 or 3 Query Request Form
- Know the approximate sample sizes needed to address the desired surveillance questions of interest using a one-time analysis and for a basic sequential design. Estimates will be made for a

range of plausible scenarios that vary the desired minimum detectable relative risk or risk difference and prevalence of use of the new product.

This assessment may be informed by data from the literature or ideally from existing data within the Sentinel environment where surveillance will be performed. The scope of this step is intentionally limited to conserve resources until more is known about the ability to conduct the assessment in Sentinel. For example, for a newly approved product, this step would likely occur in the first 6 months after approval before uptake begins or it would begin during the peri-approval period. More time-intensive planning steps would not be recommended until use of the product increases and it becomes apparent that there is enough uptake to warrant initiation of a formal safety assessment. See **Table 6** for a summary of design considerations made at this step (e.g., defining the population, outcome, exposures, key confounders, and approximate sample size needs). See **Table 7** for a detailed checklist of feasibility questions to ask and proposed data summaries to answer them. This step is designed to lay the groundwork for more detailed planning as soon as there is adequate uptake of the product of interest to perform a safety assessment.

Step 2: Conduct a descriptive *Uptake Assessment* to quantify product of interest uptake within Sentinel and describe this population. This step does *not* involve assessing product-outcome relationships. After completing this step, the surveillance team will be able to determine if:

- For an existing product of interest
  - There are an adequate number of users for a well-powered, one-time analysis
  - There are not enough users for a one-time analysis but there is adequate uptake to support the initiation of routine sequential surveillance.
  - Continued uptake monitoring is needed before initiating a safety assessment, one-time or sequential.
- For a newly marketed product of interest
  - There are an adequate number of users to support the initiation of routine sequential surveillance (or potentially even conduct a well-powered one-time assessment, if new product uptake is extremely rapid).
  - Continued uptake monitoring is needed before initiating sequential surveillance.

Initiation of detailed plans for a one-time analysis may be feasible right away if the Sentinel database contains close to (e.g., >75% of) the total sample size estimated in step 1. If uptake levels are moderate (e.g., 25-75% of the total estimated sample size from step 1), initiation of detailed plans for sequential surveillance may be feasible. If uptake levels are low (e.g., <25% of the total estimated sample size from step 1), then continued uptake monitoring should be done before further safety assessment planning occurs. In particular, a single, consolidated *Interim Uptake Assessment* could be produced on a quarterly basis containing uptake description for all products that are currently of interest to FDA for potential safety evaluation within Sentinel. See **Table 8** for a checklist of uptake assessment questions, data summaries to answer them, and rationale.

Step 3: Finalize the surveillance plan as soon as product use is prevalent enough for either a one-time or sequential evaluation. Upon completion of step 3, the surveillance team will be able to:

- Finalize entries for the required fields on the Level 2 or 3 Query Request Form
- Finalize sample size requirements needed to address the desired surveillance questions

This step involves rerunning the *Feasibility Assessment* in step 1 and using it to refine the surveillance plan based on the most current available data. In addition, more time-intensive planning steps are now

recommended to be done. This includes selecting and finalizing confounder definitions, selecting an appropriate sequential design, and finalizing sample size estimation given this design and current uptake levels. This heavier commitment of planning resources occurs only once it is evident that uptake is adequate to conduct the evaluation.

Step 4: Implement surveillance plan using the final specifications made in step 3. Upon completion of step 4, the surveillance team will be able to:

- Produce a *Final Safety Evaluation Report* if uptake was enough for a one-time analysis
- Produce an *Interim Safety Evaluation Report* if uptake was inadequate to perform a well-powered single analysis but enough to initiate sequential surveillance

Although a surveillance plan could be formulated in the absence of feasibility and uptake data, many important practical questions can be asked and answered ahead of time that will make surveillance planning more efficient and the resulting plan stronger and subject to fewer downstream modifications. As outlined above in the 4-step suggested framework, ideally surveillance plan development and feasibility assessments are performed iteratively, with feasibility data informing an initial plan and then updated information informing refinements to the final plan.

If uptake was adequate to support a one-time analysis, then the *Final Safety Evaluation Report* can be produced. If uptake was inadequate to perform a well-powered single analysis but enough to initiate sequential surveillance, then an *Interim Safety Evaluation Report* can be made. If uptake was minimal, then the *Feasibility Assessment* and the *Uptake Assessment* can be provided to illustrate the lack of uptake and describe the population of users.

Table 6. Surveillance planning questions and design considerations for cohort designs

Feasibility question	Specific considerations
How should cohort **eligibility** be defined?	Data Partners (e.g., formulary, partner size) Surveillance time frame of interest Wash-in period for enrollment Population characteristics (e.g., age, indication, comorbidity, disease severity, prior medication use, prior AEs) for inclusion/exclusion
What are the expected characteristics of those **exposed**?	Type of exposure: short/acute vs long/chronic Indications: disease(s), 1st line therapy vs other, treatment vs. prevention, off-label use Contraindication(s) Publicity/advertising Projections on market share from previous products or prevalence of indication(s)
What (active) **comparator**(s) is most appropriate and available for use?	Similar indication(s), contraindication(s), co-morbidities, and therapeutic class Prevalence of use in U.S. population
How should use of exposure and comparator(s) be defined?	Define new use or prevalent use Define continuous episodes of use if relevant What to do with medication switchers What to do with discontinuers Whether to incorporate stockpiling algorithm

Feasibility question	Specific considerations
	Window of risk (e.g., first 30 days)
Characteristics of and how to define adverse event (**AE**)?	Adopt standard validated algorithms from the Sentinel consideration of outcomes library or literature to define AE or develop (and possibly validate) new algorithms Background incidence and prevalence rates of AE Expected time to AE and censoring (death, disenrollment) Can AE be adequately ascertained from MSCDM Relevance of AE severity and ability to determine severity from MSCDM RR or risk difference (RD) of interest to detect Setting (outpatient, ED, hospital) for AE detection
What potential **confounders** should be captured?	Start with core confounder list created by Protocol Core Expand list to include covariates likely associated with exposure and AE (e.g., history of the AE) Availability of data on confounders across Data Partners Potential for residual or unmeasured confounding Classification of confounders (e.g., 10 year age groups; # of hospitalizations in prior year: 0, 1, 2+)
What **method of confounding** control should be used? [Methods considered in this report are Regression and IPTW]	# of expected confounders # that can be reasonably included Need to achieve de-identified aggregate data (for regression) Preference to use a propensity score versus standard adjustment Interest in RR (regression) versus RD (IPTW)
What **subgroup analyses** should be planned?	By confounders By indication or other comorbidity By disease severity By age By insurance type (e.g., Commercial, Medicare, Medicaid)
What **sensitivity analyses** should be planned?	Varying definition of new user, intent to treat vs standard per protocol analysis, varying definition of AE
Should a one-time or sequential analysis be conducted? If sequential, what **sequential design** should be used?	Estimate maximum sample size requirements for a one-time analysis and a simple sequential design: • Testing frequency (4-8 tests equally spaced based on # of expected AEs – i.e., in information time) • Shape of signaling threshold over time (constant) • Hypothesis (1-sided)

Table 7. Checklist of surveillance planning questions, data summaries* to answer them, and rationale

Feasibility question	Data Needed (Source: CIDA and/or Literature)	Rationale
How should the active **comparator** group be defined?	• Counts (%) of comparators overall and by Data Partner • Counts (%) of comparators by any debatable eligibility criteria	To assess availability and comparability of the comparator group. To inform sample size

Feasibility question	Data Needed (Source: CIDA and/or Literature)	Rationale
	• Median duration of use	estimation.
What is the **baseline incidence of AEs** in comparator group?	• AE counts and rates overall, by month/year for past X years, and by Data Partner • Average time to AE • AE counts and rates by key potential confounders (age, comorbidity, etc.) • AE counts and rates by severity, if relevant & possible • AE counts and rates by any debatable eligibility criteria	To inform sample size estimation.
What is the **distribution of key confounders** in comparator group (the likely surveillance population)	• Counts (%) of comparators by key confounders (e.g., age, comorbidity), overall and by Data Partner • Counts (%) of comparators jointly by key confounders and AE status	To explore prevalence of potential confounders in the population. To explore level of de-identification achievable using key confounders.

*For a more streamlined report, only the underlined summaries could be produced.

Table 8. Checklist of interim uptake questions, data summaries to answer them, and rationale

Feasibility Question	Data Needed (Source: CIDA)	Rationale
How much **new product and comparator uptake** has occurred by the end of each quarter?	• Counts (%) of cumulative new product users, overall and by month and Data Partner • Counts (%) of cumulative comparators, overall and by month and Data Partner	To inform when to begin surveillance and to conduct each planned sequential test To assess the extent to which the new product is replacing the old and there are enough new and old users to compare over time
How many **AEs among new product users and comparators** have occurred by the end of each quarter?	• AE counts and rates overall, by month, by Data Partner (and by key confounders?) • Average time to AE • Counts (%) of diagnosis source, diagnosis code, and severity if relevant & possible	To descriptively monitor trends in the numbers and rates of AE(s) To determine when there are enough AEs to begin surveillance and when to include a Data Partner To update baseline AE incidence for sample size estimation
What is the **distribution of key confounders**?	• Counts (%) of key confounders overall and by Data Partner • Counts (%) of key confounders by exposure groups • Counts (%) of key confounders by AE status	To assess extent of potential confounding. To explore level of de-identification achievable based on key confounders

*For a more streamlined report, only the underlined summaries could be produced.

D. DETAILED PLANNING STEPS TO FINALIZE SELECTED CONFOUNDERS

1. Overview

All routine evaluations, regardless of the chosen confounder adjustment strategy (e.g., matching, regression or IPTW), must determine what specific potential confounders and how many confounders should be included in the multivariable procedure. The goal is *not* to replicate the detailed and time-consuming confounder selection process that is typically performed when designing a customized safety surveillance protocol. Rather, routine evaluations intend to streamline the confounder selection process to be efficient and to include the key covariates that are expected to be the primary sources of confounding. The goal is a parsimonious model that includes only important confounders determined a priori. This is because routine surveillance activities are expected to be carried out relatively quickly and for multiple product-outcome pairs simultaneously.

The criteria for a confounding factor are: 1) It must be a risk factor (or protective factor) for the outcome of interest; 2) It must be associated with the exposure of interest. For example, diabetes severity and duration are confounders in a study of saxagliptin medication use and myocardial infarction (MI) risk because greater diabetes severity and duration are risk factors for MI, and diabetes severity and duration are likely unevenly distributed among saxagliptin users vs. users of a comparator diabetes medication; and 3) It must not be an intermediate step in the causal pathway between the exposure and outcome. For example, hypertension might not be a confounder in a study with heart disease as the outcome because hypertension might be part of the causal pathway towards heart disease.

Each routine monitoring tool faces different challenges related to confounder selection. For example, for methods that use propensity scores, there must be an adequate amount of initial product uptake at a given Data Partner to successfully build and fit a propensity score model that can accommodate the desired number of confounders. While propensity scores can effectively reduce sparse data issues in standard regression (see below) by reducing a large number of covariates to a single dimension (i.e., 1 variable), sparse confounders can also produce extreme tails in the propensity score. Regression on individual covariates has different challenges. It relies on an outcome regression analysis of data that are aggregated by exposure, outcome, and confounder strata as depicted in the following simplified example table, with one row per stratum:

Site	Age Cat	Sex	Drug	Cases	N
A	25-30	F	NME	10	1000
A	25-30	F	COMP	15	600
A	25-30	M	NME	4	2000
A	25-30	M	COMP	10	1000
…					
A	60-65	M	COMP	55	5000
B	25-30	F	NME	25	10000
B	25-30	F	COMP	88	8000
…					

The main confounder selection challenge that arises in this setting is the potential for sparse cell counts. This problem occurs when either of the following is present:

- Low cell counts (e.g., rare outcome, rare exposure, and/or rare covariate(s)) in any one particular stratum
- Too many confounder categories generating too many total strata, such that the total number of populated strata is not substantively smaller than the total number of eligible individuals

Either issue could lead to a dataset that is not sufficiently de-identified. In the next section, we detail a general strategy for confounder selection for Regression that is designed to deal with these challenges.

2. General strategy

- **Start with the core confounder list**. Include the core potential confounders proposed by the Protocol Core based on expert opinion. These covariates are referenced in the User's Guide as confounders that one would likely control for in each routine evaluation, and they are shown in the table below.

Demographics
Age
Sex
Calendar time*

Data Partner*
One or two healthcare utilization measures in baseline period**
# of visits to emergency departments
# of ambulatory visits
# of hospitalizations
# of distinct drugs ordered/dispensed
# of prescriptions ordered/dispensed
Lifestyle Factors (if relevant based on cohort age)
Smoking***, per algorithm developed by the "15 Cohorts" workgroup
Body mass index**, if available in the common data model; otherwise, per algorithm developed by the "15 Cohorts" workgroup
Combined Charlson-Elixhauser comorbidity index

*Covariate requiring special consideration given the sequential nature of planned analyses
**Commonly defined as 182 or 365 days prior to index date. Subjects must be enrolled during this assessment period. Measures are likely collinear and thus we recommend adjusting for only 1 or 2 at the maximum.
***Discretely-captured data field not currently in the Mini Sentinel Common Data Model, therefore an alternate diagnosis-based algorithm is suggested

- **Remove core confounders that are not relevant**. Inclusion of each above-listed covariate in a multivariable model should be carefully considered given the medical product-health outcome of interest (HOI) pair being examined, as not all core confounders may be true confounders for a particular evaluation. Review of the literature and/or a pre-launch CIDA run may be useful in informing this decision.
- **Remove collinear variables.** Collinearity between confounders may be especially important when the goal is to estimate the independent effects of covariates, stratify risk by covariates, or create a parsimonious standard regression model that meets aggregate data standards. Many of the health care utilization variables noted in the core confounder list above are correlated. As such, it is often appropriate to choose only 1 or 2 of these health care utilization measures as covariates. A combined comorbidity index may also be collinear with individual comorbidity flags and vice-versa. Collinearity of variables can results in changes in the coefficients of the collinear variables, inflation of the standard errors of the variables, and wide confidence intervals for the risk estimates of such variables. Coefficients for collinear covariates may not accurately estimate previously known magnitudes of risk.
- **Remove confounders not available at all participating Data Partners.** Only covariates that can be obtained at all participating Data Partners should be included. For example, smoking status is only available from electronic health record (EHR) data and thus this variable is commonly excluded as a covariate when Data Partners without EHR data are included. The MSOC can assist with information on data availability by partner. If limiting confounders creates a major limitation, consider limiting the study to only Data Partners with all covariates of interest.
- **Define categories for core confounders**. In order to aggregate data by confounder strata, categorical confounder categories must first be chosen for continuous covariates (e.g., age, # of

visits for all utilization variables, # of medications or prescriptions, and comorbidity index). Categories may be determined empirically (pre CIDA run) or based on prior protocols, the literature, and clinical relevance (e.g., higher risk in postmenopausal women).

- **Add key additional pair-specific confounders as needed.** Based on content area expertise, a brief literature review including clinical trial data, and existing experience within FDA (e.g., from pre-licensure trials or other sources), add other major risk factors for the health outcome of interest that could also be reasonably expected to be associated with the product of interest. If published studies exist, consider whether replication of the study (i.e., similar confounders) is warranted and feasible.

3. Other Issues

- Definition of 'de-identified' data. Currently, the definitions are Data Partner-specific and inconsistent. In the future, it would be advantageous to agree upon and establish well-defined criteria as to an adequate level of de-identification across Data Partners.
- Level of de-identification achieved. The data pull process may be inherently iterative since it is not possible to know in advance what distribution of confounders exists at a given Data Partner and what level of de-identification will be achieved based on identified confounders of interest and information needed on the confounders. Feasibility data should be pulled to assess this in advance to the extent possible.
- PS methods do not get an estimate of the effect of individual risk factors on the outcome. One way to regain this level of detail for methods that use a PS is to have some key individual variables kept outside the PS and adjusted separately (e.g., age, gender). This can be a reasonable approach and should be planned for during covariate selection if information about effects of important individual variables on AEs is needed.

E. DETAILED STEPS TO SELECT AND FINALIZE THE SEQUENTIAL DESIGN

In this section, we provide further detail on steps for selecting and finalizing the sequential design that were outlined in the overall suggested surveillance planning framework in Section IVC. Sequential design planning tasks occur in step 1 (Conduct a descriptive quantitative Feasibility Assessment) and step 3 (Finalize the surveillance plan) of the overall suggested framework described previously. We use the example product-outcome pair of ACE inhibitors and angioedema to illustrate these more detailed sequential design planning steps. Although this report focuses on the sequential design processes for the Regression and IPTW methods, the basic principles also apply to other approaches, such as a cohort design that involves a PS-matched cohort analysis.

In the descriptive Feasibility Assessment (step 1), which is proposed to facilitate the development of a preliminary surveillance plan, an important first step relating sequential design planning is to understand what approximate sample sizes are needed to address the identified surveillance questions of interest. Knowing this will help guide decisions about whether a well-powered one-time analysis is possible to initiate right away –OR- whether there are not enough users for a one-time analysis but there is adequate uptake to support the initiation of routine sequential surveillance –OR- whether continued uptake monitoring is needed before initiating a safety assessment, one-time or sequential.

To accomplish this, we recommend that sample sizes be estimated for both a one-time analysis and for a simple sequential design, such as one with the following specifications:

- Testing frequency: 4, 8 or 16 tests, equally spaced based on the # of expected outcomes
- Shape of signaling threshold over time: flat (constant)
- Hypothesis: 1-sided
- Power: 0.80 or 0.90
- Type 1 error: 0.05

In addition, estimating the required sample size for a one-time analysis that assumes some plausible amount of confounding is recommended, in order to assess the magnitude with which the sample size for any design (one-time or sequential) may need to be inflated due to confounding factors. Other needed inputs for sample estimation include:

- Estimated rate of the outcome among the comparator group (from the *Feasibility Assessment*)
- Estimated average follow-up time per subject (from the *Feasibility Assessment*)
- Range of the potential prevalence of the new product (e.g., 10%, 25%, 50%)
- Range of the minimum effect size of interest to detect and/or based on prior data (e.g., RR= 2, 5)

Given these inputs, all these computations can be done using available SAS statistical software version 9.3 (SAS Institute Inc., Cary, NC). See also Appendix 1 in Section VII for Code to Compute Sample Size. In particular, for each design, sample sizes can be estimated for a small range of plausible scenarios that vary the desired minimum detectable relative risk (RR) (or odds ratio [OR]) or risk difference and prevalence of use of the new product**. Table 9** displays this type of preliminary sample size data for a Regression analysis estimating the relationship between ACE inhibitors and the occurrence of angioedema within 30-days of exposure, an example that is described further in Section V (p. 33). It shows the maximum required sample size estimates for designs that vary the total number of planned analyses and across a range of assumed values for exposure uptake (i.e., the percentage of the population using ACE inhibitors versus the comparator) and minimum detectable RRs of interest. Maximum sample size is defined as the number of new ACE inhibitor users that are required to achieve 90% power to detect a specified minimum RR of interest if no signal is detected during the course of a sequential evaluation. **Table 10** shows analogous data for an IPTW analysis using a risk difference signaling criterion. Note that the range of risk differences presented in Table 10 was selected to correspond to the range of RRs presented in Table 9, for comparability.

Table 9. Maximum sample size* for Regression method by number of analyses, ACE inhibitors

% of cohort using ACE inhibitors	RR	Maximum Sample Sizes			
		1-Time	4-Times	8-Times	16-Times
25%	1.5	902,285	1,084,340	1,153,941	1,213,358
	2	**308,745**	**371,041**	**394,857**	**415,189**
	3	122,903	147,701	157,182	165,275
50%	1.5	676,714	813,255	865,456	910,019
	2	231,559	278,281	296,143	311,392
	3	92,178	110,776	117,887	123,957

*Assumptions:
Binary outcome: Angioedema in 30 days after exposure
Comparator group: Beta blockers
Estimated rate of outcome among comparator group: 3.08/10,000 person-months
Boundary shape: Flat on standardized Z-statistic scale
Power: 90% to detect a given RR

Table 10. Maximum sample sizes* for IPTW method by number of analyses, ACE inhibitors

% of cohort using ACE inhibitors	RD (per 10k person-months)	Maximum Sample Sizes			
		1-Time	4-Times	8-Times	16-Times
25%	1.5	625,032	751,145	799,360	840,519
	3	**156,258**	**187,787**	**199,840**	**210,130**
	6	39,065	46,947	49,960	52,533
50%	1.5	468,774	563,359	599,520	630,389
	3	117,194	140,840	149,880	157,598
	6	29,299	35,210	37,470	39,400

*Assumptions:
Binary outcome: Angioedema in 30 days after exposure
Comparator group: Beta blockers
Estimated rate of outcome among comparator group: 3.08/10,000 person-months
Boundary shape: Flat on standardized Z-statistic scale
Power: 90% to detect a given risk difference

Once it is evident based on the estimated preliminary sample size needs (from step 1) and the available uptake numbers (from step 2's Uptake Assessment) that product uptake is strong enough to initiate either a one-time or sequential evaluation, one can then proceed to step 3: Finalize the surveillance plan. Recall that this step involves a heavier commitment of planning resources, and so it occurs only once it is evident that uptake is adequate to conduct the evaluation. For sequential design planning, the goal of this step is to examine the properties of several potential designs in more detail so that they are fully understood prior to implementation and make a final design selection. This process should involve clear communication and collaborative vetting of several potentially suitable design(s) with FDA to assess acceptability of the design to address the primary safety aims. Sequential design choices to make, which should guide the specific data and scenarios that are examined for acceptability, include:

- Number and timing of analyses
- Shape of the signaling boundary over time
- Whether a relative or absolute measure (e.g., relative risk or risk difference) is of primary importance to monitor
- Desired magnitude for the minimum detectable difference that will define a signal
- Preferred power and Type 1 error levels

In addition, each analytic tool requires that a minimum number of adverse events be observed at each Data Partner before data from that partner can contribute to the analyses. For regression methods, where a ratio measure (e.g., an odds ratio for logistic regression or a relative risk for Poisson regression) is estimated, one event is needed in *both* the exposed and the comparator group. For IPTW, which computes a risk difference, only one event in total is needed (i.e., an event in *either* the exposed or the comparator group).

Once the final design is selected, then final estimated sample size requirements can be computed given this design. **Table 11** displays example sample size estimates for a more in-depth variety of potential sequential designs that implement a Regression analysis to estimate the relationship between ACE inhibitors and the occurrence of angioedema within 30-days of exposure. Specifically, it shows the maximum required sample sizes for designs that vary both the number of total planned analyses as well as the shape of the signaling threshold over time across a range of assumed values for exposure uptake and minimum detectable RRs of interest. All of the three signaling thresholds shown in Table 11 (flat,

O'Brien-Fleming, and an intermediate threshold that is 'in between') are part of a class known as the power family, where each has a different value for a parameter that governs the specific shape of the signaling threshold over time.[48] **Table 12** shows analogous data for an IPTW analysis using a risk difference signaling criterion.

Table 11. Maximum sample sizes* for Regression method by boundary shape

# Looks	% ACE	RR	Maximum Sample Sizes		
			Pocock	In-Between	O'Brien-Fleming
4	25%	1.5	1,084,340	971,877	931,246
		2	**371,041**	**332,558**	**318,655**
		3	147,701	132,383	126,848
	50%	1.5	813,255	728,908	698,435
		2	278,281	249,419	238,992
		3	110,776	99,287	95,136
8	25%	1.5	1,153,941	990,736	943,715
		2	**394,857**	**339,012**	**322,922**
		3	157,182	134,951	128,546
	50%	1.5	865,456	743,052	707,786
		2	296,143	254,259	242,191
		3	117,887	101,214	96,410
16	25%	1.5	1,213,358	1,003,258	951,930
		2	**415,189**	**343,296**	**325,733**
		3	165,275	136,657	129,666
	50%	1.5	910,019	752,444	713,948
		2	311,392	257,472	244,300
		3	123,957	102,493	97,249

*Assumptions:
Binary outcome: Angioedema in 30 days after exposure
Comparator group: Beta blockers
Estimated rate of outcome among comparator group: 3.08/10,000 person-months
Power: 90% to detect a given RR

Table 12. Maximum sample sizes* for IPTW method by boundary shape

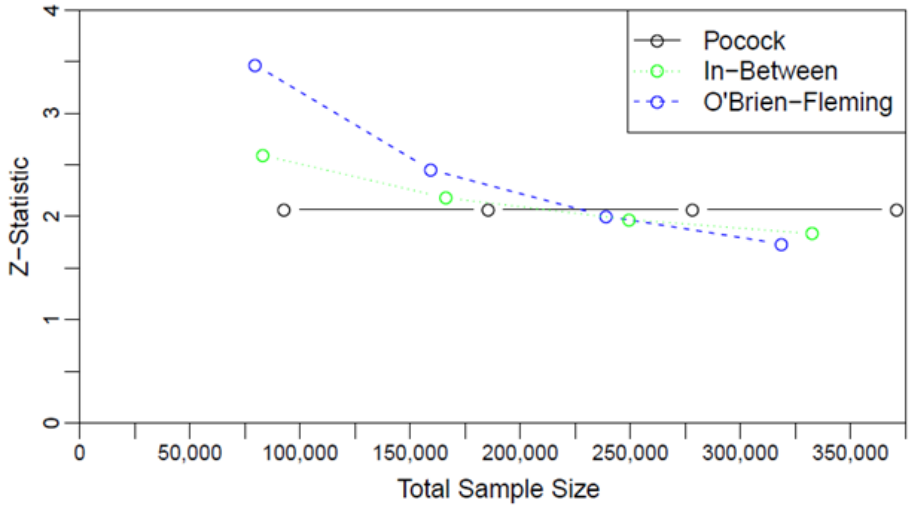
# Looks	% ACE	RD (per 10k person-months)	Maximum Sample Sizes		
			Pocock	In-Between	O'Brien-Fleming
4	25%	1.5	751,145	673,240	645,094
		3	**187,787**	**168,310**	**161,274**
		6	46,947	42,078	40,319
	50%	1.5	563,359	504,930	483,821
		3	140,840	126,233	120,956
		6	35,210	31,559	30,239
8	25%	1.5	799,360	686,304	653,731
		3	**199,840**	**171,576**	**163,433**
		6	49,960	42,894	40,859
	50%	1.5	599,520	514,728	490,298

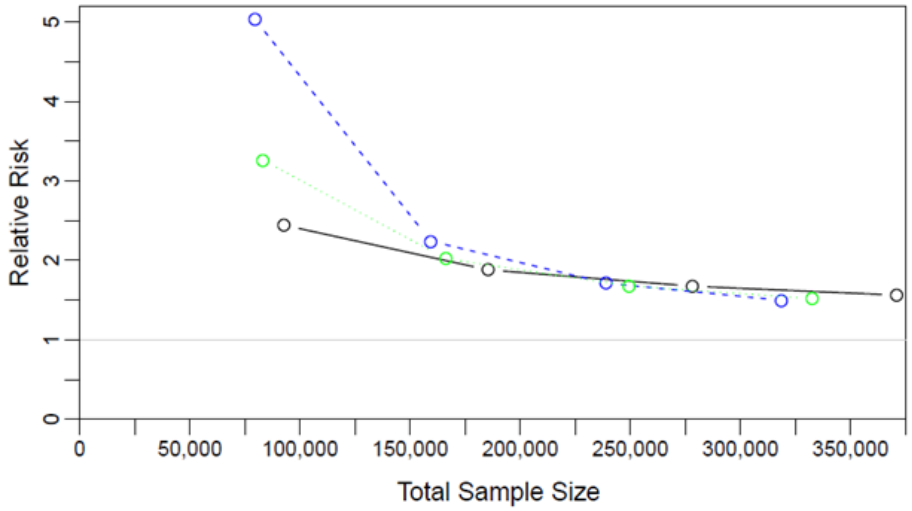
# Looks	% ACE	RD (per 10k person-months)	Maximum Sample Sizes		
			Pocock	In-Between	O'Brien-Fleming
		3	149,880	128,682	122,575
		6	37,470	32,171	30,644
16	25%	1.5	840,519	694,978	659,422
		3	**210,130**	**173,745**	**164,856**
		6	52,533	43,437	41,214
	50%	1.5	630,389	521,234	494,567
		3	157,598	130,309	123,642
		6	39,400	32,578	30,911

*Assumptions:
Binary outcome: Angioedema in 30 days after exposure
Comparator group: Beta blockers
Estimated rate of outcome among comparator group: 3.08/10,000 person-months
Power: 90% to detect a given risk difference

Figures 1 and 2 show the magnitude of the signaling thresholds over time for the **bolded** design scenarios in **Tables 11 and 12** for the Regression and IPTW methods, respectively. The top panel in each figure shows the signaling criteria on the scale of the standardized test statistic, and the bottom panel in each figure shows the signaling criteria on more interpretable scale of the risk measure of interest (i.e., on the scale of the RR for the Regression method and on the scale of the risk difference for the IPTW method). In addition, these plots depict the total number of adverse events among new users of ACE inhibitors and among new users of comparators that would generate a signal at each planned analysis time point.

Figure 1. Z-statistic and relative risk signaling thresholds for a sequential design with 4 analyses, 90% power for RR= 2

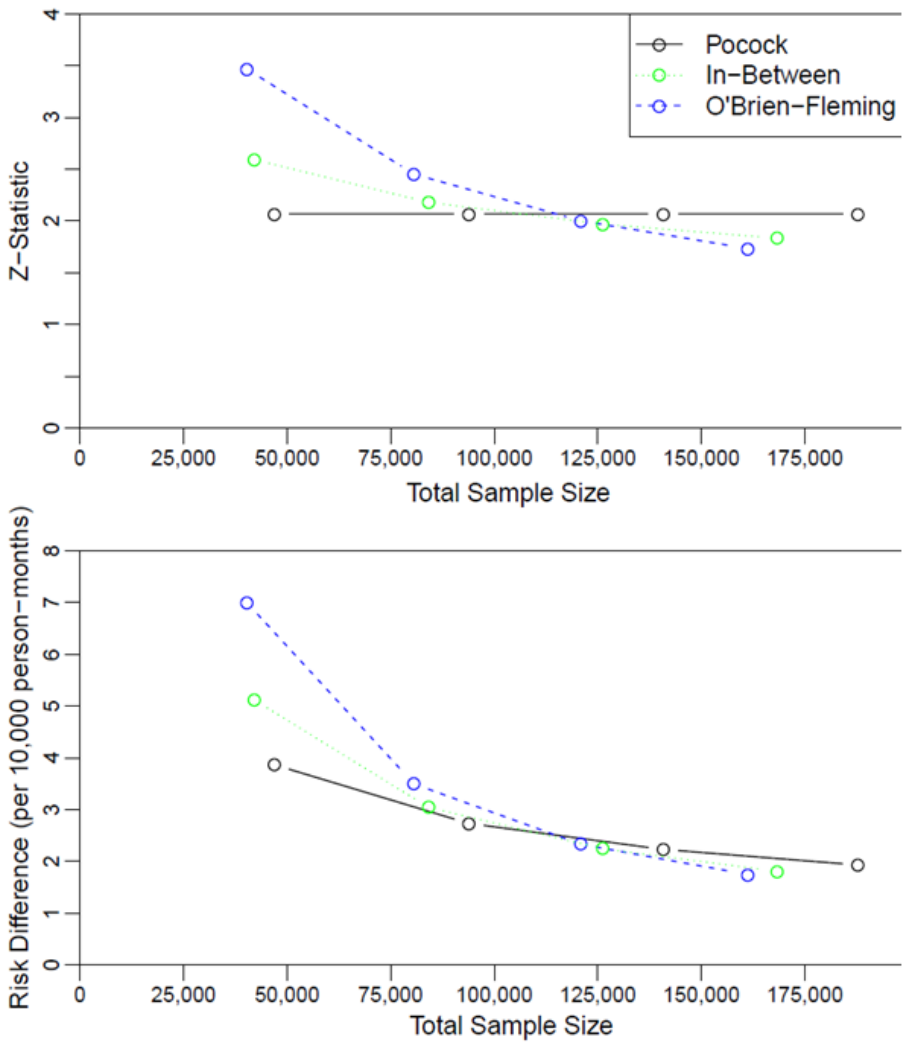




Analysis Times: Total Sample Size (Outcomes in Unexposed, Exposed)

	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Pocock:	92760(21, 17)	185520(43, 27)	278281(64, 36)	371041(86, 45)
In−Between:	83140(19, 21)	166279(38, 26)	249418(58, 32)	332558(77, 39)
O'Brien−Fleming:	79664(18, 31)	159328(37, 28)	238991(55, 32)	318655(74, 37)

Figure 2. Z-statistic and risk difference signaling thresholds for sequential design with 4 analyses, 90% power for RD=3



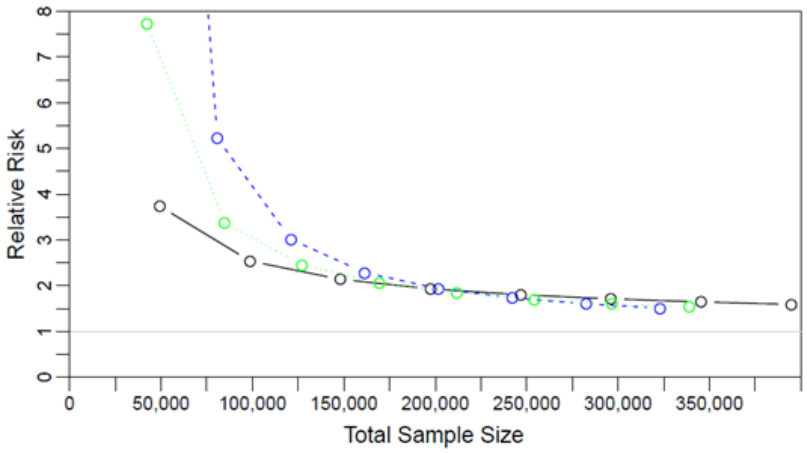
Analysis Times: Total Sample Size (Outcomes in Unexposed, Exposed)

	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Pocock:	46947(11, 8)	93894(22, 14)	140840(33, 19)	187787(43, 24)
In−Between:	42078(10, 10)	84155(19, 14)	126232(29, 19)	168310(39, 23)
O'Brien−Fleming:	40318(9, 12)	80637(19, 15)	120956(28, 19)	161274(37, 23)

This statistical information can help facilitate a dialogue among stakeholders and lead to more informed final decisions about the preferred sequential design, given the relevant scientific and practical considerations for the safety question of interest. For instance, focusing on the Regression approach, based on the data in **Table 11 and Figure 1**, a sequential design with only 4 analyses would result in the first analysis not being conducted until over 75,000 patients have been observed under any design. This may be viewed as waiting too long, if there truly is a perceived increased harm in the population. Focus might then turn to the designs with more frequent analyses, such as those with 8 or 16 total planned analyses presented in **Figure 3**. At the first analysis, all designs with 16 total tests would signal based on

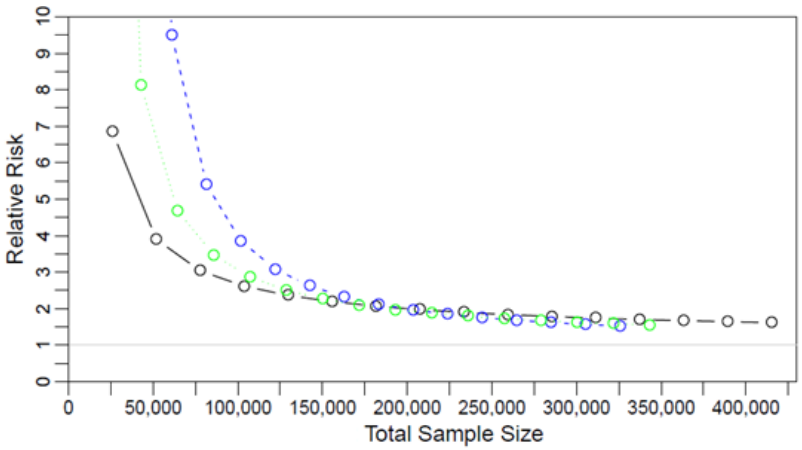
about 5 adverse events in the comparator group (**Figure 3**, bottom plot). This number of events may be deemed too small upon which to base a safety signal, which may direct further attention to the 8-analysis designs. These all require about 10 events in each group before a signal would be raised (**Figure 3**, top plot). Among those designs, the O'Brien-Fleming threshold [49] may be considered too conservative at the first analyses, requiring an extremely high RR (greater than 27, data point not shown) to generate a signal. This might lead a surveillance team to choose an 8-analysis plan with either the Pocock threshold [50] (which would signal if the RR is ~4 or more at the first analysis) or a power family threshold 'in between' these two extremes[51,52] (which would require a RR of ~8 or higher to signal at the first analysis). Similar plots for IPTW method are shown in **Figure 4**.

Figure 3. Relative risk signaling thresholds for design with 8 (top) or 16 (bottom) analyses, 90% power for RR= 2



Analysis Times: Total Sample Size (Outcomes in Unexposed, Exposed)

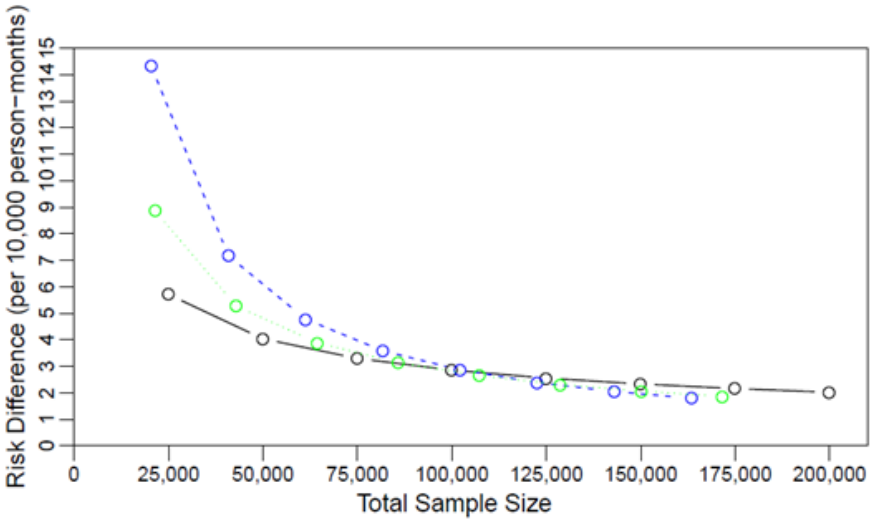
	Analysis 1	Analysis 2	...	Analysis 8
Pocock:	49357(11, 14)	98714(23, 19)		394857(91, 48)
In-Between:	42376(10, 12)	84753(20, 17)		339012(78, 42)
O'Brien-Fleming:	40365(9, 12)	80730(19, 16)		322922(75, 40)



Analysis Times: Total Sample Size (Outcomes in Unexposed, Exposed)

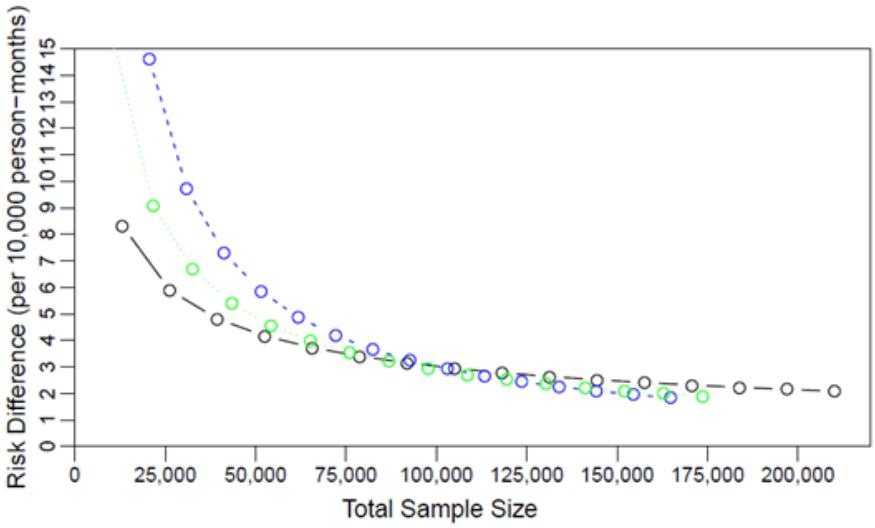
	Analysis 1	Analysis 2	...	Analysis 16
Pocock:	25949(6, 14)	51899(12, 16)		415189(96, 52)
In-Between:	21456(5, 11)	42912(10, 13)		343296(79, 43)
O'Brien-Fleming:	20358(5, 11)	40717(9, 12)		325733(75, 41)

Figure 4. Risk difference signaling thresholds for design with 8 (top) or 16 (bottom) analyses, 90% power for RD = 3



Analysis Times: Total Sample Size (Outcomes in Unexposed, Exposed)

	Analysis 1	Analysis 2	...	Analysis 8
Pocock:	24980(6, 5)	49960(12, 9)		199840(46, 25)
In−Between:	21447(5, 7)	42894(10, 10)		171576(40, 25)
O'Brien−Fleming:	20429(5, 11)	40858(9, 13)		163433(38, 24)



Analysis Times: Total Sample Size (Outcomes in Unexposed, Exposed)

	Analysis 1	Analysis 2	...	Analysis 16
Pocock:	13133(3, 4)	26266(6, 6)		210130(49, 27)
In−Between:	10859(3, 6)	21718(5, 8)		173745(40, 26)
O'Brien−Fleming:	10304(2, 11)	20607(5, 12)		164856(38, 26)

These sequential design and sample size planning steps can be implemented for any of the available multivariable risk estimation and testing tools that have been developed for Sentinel. As described in

this section, sample size estimation involves selecting the frequency of interim sequential testing and determining the magnitude of the critical value that defines a signal at each interim analysis. Different design choices yield different trade-offs between the probability and timing of true and false positive signals, and thus final planning decisions should be informed by a systematic evaluation of these performance characteristics as demonstrated in this section.

F. SUMMARY

The ACE inhibitors and angioedema and the ARBs and angioedema examples in Section V illustrate the type of statistical information that could be used to communicate the operating characteristics of different sequential designs to stakeholders prior to surveillance implementation. And, in an over-simplified way, it shows how such information could be used to compare the performance of competing designs, facilitate a dialogue among stakeholders about their design preferences, and lead to more informed final decisions about the choice of appropriate signaling thresholds. Clearly, though, the factors that influence the choice of sequential design selection are more complicated than this illustration conveys. Numerous scientific, ethical, and practical considerations (e.g., the magnitude of the vaccine or drug's benefit, the prevalence and severity of the adverse event of interest, etc.) should bear on this choice, and the relative importance of each factor may depend on the specific safety question of interest. Our intent here is not to comprehensively discuss these factors but rather to describe a high-level suggested framework for how statistical information can be used by stakeholders to better weigh these factors when making sequential design decisions.

The ACE and angioedema and the ARBs and angioedema are examples only. The suggestions on sequential surveillance planning within Sentinel was developed concurrently with our evaluation of the ACE/ARB examples, which are not new drugs, and thus we did not exactly follow our planning tools for a feasibility assessment that would be necessary when evaluating a new drug. However, the study by Toh and colleagues was used for assessing feasibility (Step 1) and uptake (Step 2).

V. EXAMPLE: RISK OF ANGIOEDEMA WITH ANGIOTENSIN-CONVERTING ENZYME INHIBITORS (ACE INHIBITORS) AND ANGIOTENSIN RECEPTOR BLOCKERS (ARB) COMPARED WITH BETA BLOCKERS (BB)

A. BACKGROUND

The sequential analyses described in this section were conducted to further test and enhance the Regression and IPTW modules. The example of angioedema associated with ACE inhibitors was chosen for our main analysis because:

- Angioedema is a known adverse effect of ACE inhibitors, so that time-to-signal detection can be compared between methods
- This is a drug example (since both the Regression and IPTW modules have been tested using a vaccine example previously)
- The risk of angioedema is highest shortly after initiating ACE inhibitors but continues with long-term ACE inhibitor use, making it suitable for implementation with either a short-term (applicable to Regression or IPTW) or longer-term (i.e., 1 year and applicable to Regression) risk window relative to exposure initiation

- ACE inhibitors are commonly-used medications with demonstrated uptake at multiple large Mini-Sentinel Data Partners, resulting in adequate power for comparisons across Regression and IPTW methods
- This drug-outcome pair was already evaluated within Mini-Sentinel in a protocol-based assessment and using a different methodology, which allowed us to leverage existing Mini-Sentinel definitions and coding[53]

We used this existing information to fully develop our Step 1 planning process. To fully illustrate the sequential capabilities of the Regression and IPTW methods, we conducted selected secondary analyses of the risk of angioedema comparing ARBs to beta-blockers. The uptake of ARBs was considerably slower than for ACE inhibitors (Step 2 planning process) and the effect size was much lower (hazard ratio of approximately 1.5 in Toh et al study).[53]

B. METHODS

1. Cohort Identification

The cohort included subjects who were 18 years and older at some point January 2003 through December 2012. This period was chosen since angiotensin receptor blockers (ARBs) came to market beginning in 2003. Subjects were members from one of four health plans selected based on operational considerations: Aetna (data available 1/1/2008-9/30/2013), United (data available 1/1/2008-9/30/2013), GH (data available since 2004) and KPNC (data available 1/1/2003-11/30/2013). Eligible subjects met all of the following inclusion criteria: (i) A new user of any of the medication therapeutic classes under study (angiotensin-converting enzyme inhibitors (ACE inhibitors) or angiotensin receptor blockers (ARBs)) or the comparator class (beta blockers (BBs)) ; the dispensing date that qualifies the subject as a new user was assigned to be the index date; (ii) A continuous enrollment at their health plan (defined as having a gap of less than 45 days between enrollment periods) of at least 183 days prior to index date; (3) Both medical and drug benefits existed at the index date and in the 183 days prior to the index date. Subjects were excluded from the cohort if they concomitantly used medications in multiple therapeutic classes of interest on the index date (i.e., filled more than 1 medication of interest on index). Subjects were also excluded if they had a prior diagnosis of angioedema defined by having an ICD-9 code of 995.1 recorded in any position during an outpatient, inpatient or emergency department encounter in the 183 days prior to the index date.

Subjects were followed from index date until the earliest occurrence of the outcome, disenrollment from the health plan, cessation of the therapeutic class, initiating another therapeutic class of interest, or the end of follow-up period.

2. Exposures

This evaluation examines the two medication classes of ACE inhibitors and ARBs. Users of BBs comprised the comparator group. Both single and combination products with non-study drugs of these medication classes taken orally were included (**Table 13)**. The cohort included new users of any of these three medication therapeutic classes. A new user was defined as having at least 1 outpatient dispensing for a medication in a particular therapeutic class during the study period of 2003-2012, with no dispensings for any medication in the three classes of interest in the prior 183 days. For each subject, only the first episode of exposure to one of the three medication classes of interest was included. Subjects were not allowed to re-enter the cohort even if they qualified as a new user of another class in subsequent years

of the study. Aliskiren (a direct renin inhibitor) was not evaluated since there was a small number of users.

Table 13. List of (oral) medications in each of the three therapeutic classes included in the evaluation

ACE inhibitors*	**ARBs**	**BBs**
o Benazepril (Lotensin) o Captopril (Capoten) o Enalapril (Vasotec) o Fosinopril (Monopril) o Lisinopril (Prinivil, Zestril) o Moexipril (Univasc) o Perindopril (Aceon) o Quinapril (Accupril) o Ramipril (Altace) o Trandolapril (Mavik)	o Azilsartan o Candesartan (Atacand) o Eprosartan (Teveten) o Irbesartan (Avapro) o Losartan (Cozaar) o Olmesartan (Benicar) o Telmisartan (Micardis) o Valsartan (Diovan)	o Acebutolol (Sectral) o Atenolol (Tenormin) o Betaxolol (Kerlone) o Bisoprolol (Zebeta) o Carteolol o Carvedilol (Coreg) o Labetalol (Trandate, Normadyne) o Metoprolol (Lopressor, Toprol) o Nadolol (Corgard) o Nebivolol (Bystolic) o Penbutolol (Levatol) o Pindolol (Visken) o Propranolol (Inderal) o Timolol (Blocadren)

* Enalapril/diltiazem was discontinued in 1999 and thus, not included in the list of ACE inhibitors.

When creating treatment episodes, a stockpiling algorithm was applied. The stockpiling algorithm accounts for the fact that members may refill their drug prescriptions before the end of the days' supply of the prior prescription. For example, if a member receives a 30-day dispensing for ACE inhibitors on January 1st and then receives a second 30-day dispensing for ACE inhibitors on January 20th, the stockpiling algorithm adjusts the second dispensing so that it starts on January 31st, after the first dispensing has been used in full. The treatment episode is therefore 60 days in total, through March 1st (assuming February has 28 days).

Similar to Toh et. al.'s study on angioedema with ACE inhibitors and ARBs, we used a 14-day episode gap for treatment episodes to account for non-compliance/late refills.[53] The episode gap is defined as the maximum interruption in supply (in days) allowable between two claims of the same query group to be considered part of the same treatment episode. If a gap in treatment between two claims of the same treatment was smaller than or equal to the episode gap, the algorithm "bridged" the two claims to build a continuous treatment episode that accounts for late refills. If, however, the gap between the two claims exceeded the episode gap, the treatment episode ended at the end of the first claim. The episode gap was assessed after claim service dates were adjusted by the stockpiling algorithm. After accounting for stockpiling, the end date for the treatment episode was defined as the adjusted date when the last prescription in the episode was considered to run out (i.e., adjusted fill date after accounting for stockpiling of last prescription + days' supply of the prescription). In contrast to Toh, et. al., we assumed no residual exposure effect for the drugs and therefore used a 0-day episode extension for the "at risk"

period instead of a 14-day extension because of our short windows of assessment (30 days) for most analyses.

3. Outcomes of interest

The primary outcome of interest was angioedema, defined by an ICD-9 code of 995.1 recorded in any position during an outpatient, inpatient or emergency department encounter that occurred between index date (i.e., new drug start) and end of follow-up. Angioedema that occurred on the index date was included in the outcome definition. Two risk windows of interest were defined and used in the analyses: 1) First occurrence of the outcome in the first 30 days of follow-up, and 2) first occurrence of the outcome in the first 360 days of follow-up.

4. Potential confounders

To determine what confounders to include, we first considered the set of standard confounders defined in the PROMPT Users' Guide , [8]and then considered the confounders included by Toh and colleagues in their evaluation of risk of angioedema with ACE, ARBs, and aliskiren. [53] We then eliminated some of the standard confounders in the Users' Guide for one of the following reasons: 1) They were not deemed likely to be associated with both the medications (ACE inhibitors, ARBs or BBs) and the outcome of angioedema, 2) They were not yet consistently available or there was suspicion of significant misclassification within the MSCDM, or 3) They were duplicative or non-specific (e.g., multiple overlapping health care utilization variables). A history of conditions similar to the outcome of interest (i.e., allergic reactions other than angioedema) was added as this was identified as a potentially strong predictor of angioedema that could also be related to use of these medications. This process (i.e., beginning with the standard confounder list and then subtracting and adding key confounders specific to the drug-outcome pair of interest based on expert consensus) is a feasible, streamlined approach that could be used in future routine surveillance activities. This process is also advantageous because it results in a relatively parsimonious set of key confounders, which eases model interpretation. In addition, parsimony is critical for implementation of the GS GEE regression when adjusting for individual confounders (versus, for example, a propensity summary score) since this implementation of the method requires that the data be aggregated by categories of the selected individual confounders. In order for such aggregation to yield sufficiently aggregated datasets that may be released by Data Partners for central analyses, the number of confounders needs to be relatively moderate. Of note, the GS GEE method can avoid this issue if regression adjustment is based on a propensity score (e.g., categories of the propensity score or a smooth function of the propensity score) instead of adjustment for individual confounders. The GS IPTW method also does *not* require a relatively moderate number of confounders since aggregated data are not involved and all confounders are summarized into a site-specific propensity score used for weighting in a site-specific regression model. Like any method that uses a propensity score, it simply requires that there be an adequate amount of exposure relative to the number of confounders so that the propensity score model can be fit stably.

The final list of potential confounders is shown in **Table 14**, and it is similar to the confounders included in the customized protocol-based assessment by Toh, et al.[53]

Table 14. List of potential confounders.

	BROAD CATEGORIES OF CONFOUNDERS	SPECIFICS
1	Demographics	• Age at index date (18-44, 45-54, 55-64 and 65+ years) • Sex

	BROAD CATEGORIES OF CONFOUNDERS	SPECIFICS
		• Index year (2003-2012) • Health plan
2	History of comorbid conditions in 183 days prior to index date	• Allergic reactions other than angioedema (0=no, 1=yes) • Diabetes (0=no, 1=yes) • Heart failure (0=no, 1=yes) • Ischemic heart disease (0=no, 1=yes) Note: these conditions will be defined using the same ICD-9 codes as in Toh, et al.[53]
3	Utilization in 183 days prior to index date	• At least 1 inpatient hospitalization (0=no, 1=yes)
4	Medication use in 183 days prior to index date	Any prescription* for • Traditional NSAIDS (0=no, 1=yes), include o Diclofenac (Cataflam, Voltaren) o Diflunisal (Dolobid) o Etodolac (Lodine) o Fenoprofen (Nalfon) o Flurbiprofen (Ansaid) o Ibuprofen (Advil, Motrin, Nuprin) o Indomethacin (Indocin) o Ketoprofen (Orudis, Oruvail) o Ketorolac (Toradol) o Meclofenamate (Meclomen) o Mefenamic Acid (Ponstel) o Meloxicam (Mobic) o Nabumetone (Relafen) o Naproxen (Naprosyn, Aleve, Anaprox) o Oxaprozin (Daypro) o Piroxicam (Feldene) o Salsalate (Disalcid) o Sulindac (Clinoril) o Tolmetin (Tolectin) • Aspirin (0=no, 1=yes) • COX-2 inhibitors (0=no, 1=yes), include: o Rofecoxib (Viox) o Celecoxib (Celebrex) o Valdecoxib (Bextra) • Oral corticosteroids (0=no, 1=yes)** o Prednisone o Prednisolone o Methylprednisolone o Triamcinolone o Hydrocortisone o Betamethasone o Budesonide o Cortisone o Dexamethasone

*Over the counter medications not included

** Paramethasone was discontinued in 1997 and thus, not included in the list of oral corticosteroids.

5. Statistical analyses

Primary and secondary analyses

Our planned primary analyses for conducting this methods evaluation were the shorter- and longer-term comparisons of the risk of angioedema associated with ACE inhibitors versus that of beta-blockers. If we were conducting true analyses, as opposed to methods evaluations, we would have conducted a feasibility assessment to determine the size of the available samples and whether one-time or group sequential designs were indicated. However, the magnitude of the signal between ACE inhibitor use and angioedema risk documented in Sentinel by Toh et al. was known to be fairly strong (hazard ratio of about 3), and the use of ACE inhibitors within the Sentinel population was known to be common. Therefore, we expected to have considerably more power than needed to detect this signal and anticipated that it could be detected quickly, perhaps without requiring multiple sequential analyses. Thus, to fully illustrate the sequential capabilities of the Regression and IPTW methods, we conducted selected secondary analyses of the risk of angioedema comparing ARBs to beta-blockers, where uptake of ARBs was considerably slower and the effect size was much lower (hazard ratio of about 1.5 in Toh et al).[53] For this second example pair, we conducted only Analyses 1 and 3 described below. Details on each of these analyses are presented in the next sections.

Sequential analyses

Using the sequential design described below we performed the following four sequential analyses to evaluate the relative risk of angioedema associated with the use of ACE inhibitors compared to BBs:

7. Analysis 1: Regression method with up to 30 days of follow-up for angioedema events
8. Analysis 2: Regression method with up to 360 days of follow-up for angioedema events
9. Analysis 3: IPTW method with up to 30 days of follow-up for angioedema events AND requiring 30 days of exposure use
10. Analysis 4: IPTW method with up to 30 days of follow-up for angioedema events AND no minimum required duration of exposure (i.e, did not require 30 days of post-exposure enrollment in order to be in the cohort)

The strengths and limitations of regression and IPTW methods are summarized in Table 1 (Section II).

Sequential Design (Step 2 and Step 3)

As new users of ACE inhibitors (primary exposure), ARBs (secondary exposure) and BBs (comparator) were observed over time within the surveillance cohort, we conducted routine sequential hypothesis tests that evaluated the association between these medications and the risk of angioedema using both the Regression and IPTW methods. Both methods tested a one-sided hypothesis that the risk of angioedema was elevated for ACE inhibitor users or for ARB users compared with a common comparator group of BB users. The following sequential design specifications were planned:

- Planned number of sequential looks: 4
- Signaling threshold: “In-Between” (or a power family threshold with a shape parameter of 0.25)

We planned to conduct four sequential analyses each time that a total of about 83,140 new users accrued into the cohort (Figure 1). In other words, four analyses were planned at total sample sizes of 83,140, 166,279, 249,418, and 332,558 new users of the exposure of interest (either ACE inhibitors or ARBs) and the comparator (BBs) combined. For the regression method, the sequential plan was designed to detect a doubling of the risk of angioedema between the exposure of interest and comparator with 90% power. For consistency, this same sequential design was used for the IPTW

analysis, resulting in additional power compared to the regression method. For both the Regression and IPTW methods, if the test statistic remained below the preset signaling threshold at any planned analysis time point, then surveillance was continued. If the test statistic exceeded the signaling threshold at any planned analysis time point, then the null hypothesis of no difference in angioedema risk between the exposure groups was rejected and surveillance ended early. In other words, fewer than the four planned analyses would be done if a signal was detected at analyses one, two or three. Surveillance ended without rejecting the null hypothesis if the fourth and final planned analysis time was reached and the test statistic did not exceed the threshold at any analysis time point. This framework accounts statistically for multiple tests and the early signaling rules that were imposed.

Analysis 1 - Regression analysis with up to 30-days of follow-up

Since angioedema risk is highest in the first 30 days of ACE inhibitor use,[53] we used the Regression method to evaluate the relative risk of angioedema compared to BB use in a 30-day follow-up period after initiation of either drug.[54] The confounding is taken into account through regression-based adjustment for covariates measured at baseline, and, in this case, that are thought to be associated with both the receipt of an ACE inhibitor or beta-blocker prescription and with the risk of angioedema. Like the PS Matching approach, the Regression method is suitable for use in cohort designs where the risk of outcome, e.g. angioedema, in users of a product of interest, e.g. ACE inhibitors, are compared to those who receive an alternative exposure, e.g. beta-blockers. Since ACE inhibitors and beta-blockers are often used over several months or years, it is appropriate to account for the duration of exposure, or time at risk, in the analysis. The Regression method accomplishes this through the inclusion of the natural log of time at risk as an offset term in the model. This can be thought of as Poisson regression using a Generalized Estimating Equation (GEE) framework, which is similar in concept to what is sometimes referred to as modified Poisson regression.

Event counts and person-time were aggregated by categorical drug exposure and confounder categories at each Data Partner and were then combined and de-aggregated at a central location to conduct the analysis. The method used to de-aggregate the data is described in Section IX (Appendix 3). As described above, the analysis was conducted in a formal sequential monitoring framework with pre-specified early signaling criteria.

As noted previously, in order to evaluate the Regression method in a context where the initial available sample size provided insufficient power to perform a one-time analysis, we used the same methodology to assess the relative short-term angioedema risk associated with ARB use compared to beta-blocker use as a secondary analysis.

Analysis 2 - Regression analysis with extended follow-up (i.e, 1 year)

In addition to evaluating the relative risk of angioedema comparing ACE inhibitor use to BB use in a 30-day risk window as described in Analysis 1, we also examined the relative risk of angioedema in a longer follow-up time after the index date. Regression method can flexibly and robustly accommodate a variety of different exposure and outcome types, i.e., both events that occur acutely after drug initiation or events that occur after longer-term follow-up and chronic use of a drug. Using the same cohort and sequential design as in Analysis 1, the Regression method was used to estimate the relative risk of angioedema comparing ACE inhibitor use with BB use but with a risk window that was expanded from 30 days following the beginning of a treatment episode to 360 days. Since we evaluated angioedema risk in the first 30-days in Analysis 1, we did not specifically include terms in the model to allow a differential rate of angioedema over time, i.e., this analysis assumes that relative risk of Angioedema

comparing ACE inhibitor use to BB use remains constant over the 360-day risk window. The risk rate will be unknown for most drugs. This is noted as a limitation in the discussion section.

Analysis 3 – IPTW analysis with 30-day follow-up and required exposure duration of 30 days

In addition to using the Regression method in Analysis 1 to assess the 30-day relative risk of angioedema, we also used the IPTW method to evaluate the difference in risk in the first 30-days between ACE inhibitor users and BB users, employing the same sequential design specifications as in Analyses 1 and 2. A key difference from Analysis 1 was that subjects in Analysis 3 were required to have a minimum of 30 days of exposure to the therapeutic class of interest with no censoring events occurring within the 30 days after the index date. Here a censoring event was defined as disenrollment, cessation of the therapeutic class, or initiating another therapeutic class of interest. This was done because the IPTW method assumes equal follow-up time for all patients, and this restriction ensures that all patients have an equal opportunity (i.e., an equal amount of time) to experience a potential outcome. This type of requirement is not necessary for the Regression method because, when implementing Poisson regression, variable amounts of exposure duration are taken into account in the modeling as described in Analysis 1 above.

The IPTW method is a flexible approach for cohort designs where a short-term exposure and an acutely occurring outcome are of interest.[55] The method performs site-stratified, IPTW estimation and group sequential testing in a distributed data setting where the quantity of interest is the overall adjusted risk difference (RD). Specifically, a Data Partner-specific propensity score model is fit using individual-level exposure and confounder data that does not leave the Data Partner site. Adjustment for confounding is achieved by fitting a weighted linear regression model where weights are based on the Data Partner-specific propensity scores. Specifically, weighting is based on the inverse of the predicted probability of exposure, which in this example corresponds to the probabilities of receiving an ACE inhibitor or BB prescription. From this model, a Data Partner-specific adjusted RD is estimated, as well as an estimated variance that correctly accounts for variability in the IPT weights. The Data Partner-specific adjusted RD and variance estimates are transferred to a central location where they are combined via Mantel-Haenzel-type methods to provide a single overall adjusted risk difference estimate with corresponding standard error. Given the overall risk difference and standard error estimates, a standardized test statistic is calculated (RD/sqrt(var(RD))) and compared to a preset signaling threshold to determine if there is an elevated risk or whether sequential monitoring should continue. One advantage of this approach is that it strongly controls for site confounding if it exists, but has also been shown to be as efficient as a non-stratified estimate when no site-level confounding is present.[11] Additionally, only one event is required to estimate a risk difference (unlike methods that use relative measures, which require one event in each exposure group), making it well suited to a rare event scenario.

As noted above, in addition to evaluating the difference in risk comparing ACE inhibitor use and BB use in the first 30-days of exposure, we also evaluated the difference in risk comparing ARB use to BB use in the first 30-days of exposure as a secondary analysis.

Analysis 4 - IPTW analysis with 30-day follow-up and no minimum exposure duration requirement

This IPTW analysis also used the same sequential design specifications as in Analyses 1-3. It was identical to Analyses 1 and 3 with respect to the risk window definition (i.e., first occurrence of angioedema within the 30-day risk window) and identical to Analysis 1 and 2 with respect to the included cohort (i.e., it did *not* restrict to subjects who had at least 30 days of exposure duration as in Analysis 3 but rather included subjects with *any* length of exposure duration). Unlike the Regression method however, the IPTW method does not account for these differences in exposed person-time when comparing

differences in risk. Rather, it assumes that all subjects were observable for the full follow-up time of 30 days. It is still reasonable to apply IPTW in this way since follow-up time is very short and so we expect very few subjects to be censored during this very short follow-up period. Thus, Analyses 3 and 4 will likely yield similar results.

Results (Step 4)

We began surveillance on June 29, 2008, when all Data Partners had data available for analysis. In the sections below, we first report our primary analysis results for all four planned analyses of ACE inhibitors versus BBs. In each of these analyses, the sequential signaling threshold was crossed at the first of the four planned looks, indicating an elevated risk of angioedema associated with ACE inhibitor use. Thus, all results presented for the ACE inhibitor reflect the data that had accumulated at the time of the first analysis.

Because all four analyses for ACE inhibitors signaled at the first look, we also present secondary results of selected analyses (i.e., Analyses 1 and 3) for a second example pair: ARB use and angioedema risk. For both of these ARB analyses, no signal was detected at any of the four planned looks. Thus, all results presented for ARBs reflect the data that had accumulated at each of the four planned analyses. For both the ACE inhibitors and the ARBs comparisons, we report on the uptake of each medication, baseline patient characteristics at the last planned analysis that was executed, overall results from each of the planned sequential analyses up to the final analysis that was reached and site-specific results from the last sequential analysis that was executed.

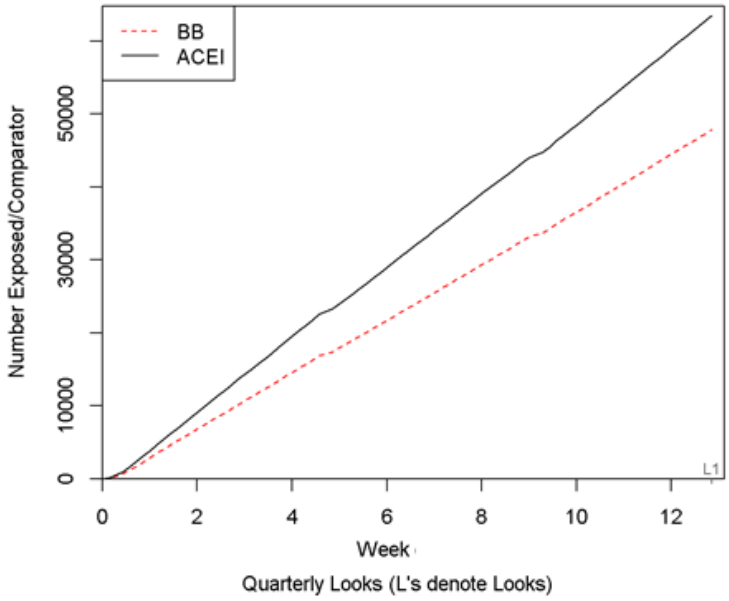
C. RESULTS

1. ACE inhibitors and risk of angioedema

The uptake of ACE inhibitors and BBs was sizeable and steady across the Data Partner sites (**Figure 5**). The use of ACE inhibitors was slightly higher than that for BBs. Since the data within Sentinel Data Partners are refreshed every quarter, the first possible time that a sequential analysis could be conducted was after the first quarter of surveillance (i.e., on September 27, 2008). By then, we had already observed more than the planned analysis 1 sample size of 83,140 total users. Thus, we conducted look 1 using data available on September 27, 2008, after one quarter of data had accrued with a total sample size of 111,329.

For the Regression method (Analysis 1 and Analysis 2), Data Partner site 4 was excluded from the analysis because there were no events in the BB group and an estimated rate ratio could therefore not be computed for that site. Subsequent regression results for ACE inhibitors and BBs therefore include sites 1, 3 and 15. For the IPTW method (Analysis 3 and Analysis 4), data were successfully returned for central analysis by sites 1, 4 and 15. An error occurred at site 3, and so no data were available from this site for this IPTW analysis. In contrast to the Regression method, however, we were able to include site 4 in the IPTW analyses despite the lack of events in the BB group because the risk difference estimate does not require that events occur in both groups. Thus, IPTW analyses for ACE inhibitors included sites 1, 4, and 15.

Figure 5: Total uptake of ACE inhibitors (ACEI) **and BBs over time at the final look**



Analysis 1 - Regression analysis with 30-day follow-up

Patient Characteristics

Comparisons of demographic and risk factor characteristics of ACE inhibitor and BB users included in Analysis 1 are presented in **Table 15**. At the first and only look (stopped due to finding of significant elevated risk), 111,329 individuals across the 3 sites were included in the analysis. Approximately 57% were ACE inhibitor users. Compared to BB users, ACE inhibitor users were more likely to be male, 45 years of age or older, and have a diagnosis of diabetes. ACE inhibitor users were less likely than BB users to have a diagnosis of ischemic heart disease or an inpatient encounter in the prior 12 months. The larger sites (1 and 3) contributed the majority of medication exposure.

Table 15: Baseline demographics of surveillance cohort members included in Analysis 1, by ACE inhibitor (ACEI) and Beta Blocker (BB) exposure at the final look, total sample size=111,329

	Total (n=111329) N (%)	BB (n=47827) N (%)	ACEI (n=63502) N (%)
Sex			
Male	56075 (50.4)	20598 (43.1)	35477 (55.9)
Female	55254 (49.6)	27229 (56.9)	28025 (44.1)
Age			
18-44	35933 (32.3)	17339 (36.3)	18594 (29.3)
45-54	33013 (29.7)	12712 (26.6)	20301 (32.0)
55-64	26975 (24.2)	10520 (22.0)	16455 (25.9)
65-99	15408 (13.8)	7256 (15.2)	8152 (12.8)
NSAIDS			
No	97464 (87.5)	41743 (87.3)	55721 (87.7)
Yes	13865 (12.5)	6084 (12.7)	7781 (12.3)
Oral Corticosteroids			

	Total (n=111329) N (%)	BB (n=47827) N (%)	ACEI (n=63502) N (%)
No	103184 (92.7)	43757 (91.5)	59427 (93.6)
Yes	8145 (7.3)	4070 (8.5)	4075 (6.4)
Allergic Reactions			
No	99418 (89.3)	42189 (88.2)	57229 (90.1)
Yes	11911 (10.7)	5638 (11.8)	6273 (9.9)
Diabetes			
No	93115 (83.6)	43362 (90.7)	49753 (78.3)
Yes	18214 (16.4)	4465 (9.3)	13749 (21.7)
Ischemic Heart Disease			
No	102021 (91.6)	41577 (86.9)	60444 (95.2)
Yes	9308 (8.4)	6250 (13.1)	3058 (4.8)
1+ Inpatient Hospital Stay			
No	99154 (89.1)	39787 (83.2)	59367 (93.5)
Yes	12175 (10.9)	8040 (16.8)	4135 (6.5)
Site			
1	46404 (41.7)	20334 (42.5)	26070 (41.1)
3	52314 (47.0)	22391 (46.8)	29923 (47.1)
15	12611 (11.3)	5102 (10.7)	7509 (11.8)

Sequential Results from the Regression Method

The first look for the 30 day analysis was after one quarter of data (i.e., 90 days) had accrued. A signal of increased risk of angioedema among ACE inhibitor users compared to BB users was identified at the first planned look (Table 16). At the time of the signal, there were 8 angioedema events observed among 47,827 BB users compared with 47 events among 63,502 ACE inhibitor users. The adjusted incidence rate of angioedema within the first 30 days of exposure in the ACE inhibitor group was 11.3 per 1000 person-years and 2.5 per 1000 person-years in the BB group. The adjusted rate ratio based on the Poisson regression comparing these groups was 4.6 indicating a considerable elevation in risk for ACE inhibitor users.

Table 16: Results of adjusted sequential monitoring using the Regression method comparing by ACE inhibitors (ACEIs) and Beta Blockers (BBs) on angioedema risk over first 30 days of use, in Analysis 1

Look	Days	BB N	BB Event (Rate)	ACEI N	ACEI Event (Rate)	BB Adj Rate	ACEI Adj Rate	Adj RR	Score Test	Boundary	Error Spent	Signal
1	90	47827	8(2.56)	63502	47(10.92)	2.46	11.30	4.60	4.71	3.08	0.000	Yes

Adjusted Poisson regression model applied using GEE framework with sequential monitoring boundaries based on permutations. Covariates included sex, age, prescription NSAIDS use, oral corticosteroid use, allergic reactions, diabetes, ischemic heart disease, any inpatient hospital stay, and site.
Abbreviations: Event (Rate)=Number(Rate per 1000 person-years) of angioedema within look and covariate category, Adj Rate=Adjusted risk (per 1000 person-years) from adjusted Poisson regression model assuming entire population was either exposed or unexposed, Adj RR=Adjusted Rate Ratio comparing ACEI to BB from Poisson regression model, Score Test=Score Test statistic from GEE Poisson regression model, and Boundary=Sequential Boundary to compare the Score test estimate.

Site-Specific Results

To descriptively explore the trends in the rate ratio estimates across the Data Partner sites, we examined results within each Data Partner at the time of the observed signal at look 1 (**Table 17**).

Adjusted event rates among ACE inhibitor users (8.2, 10.9, and 24.8 per 1,000 person-years) and BB users (0.7, 2.6, and 8.3 per 1,000 person-years) varied greatly across sites but were consistently higher among ACE inhibitor users within each site. The variation observed across sites is not surprising given the small number of events within each site (e.g., <5 at each site in the BB group). The site-specific adjusted rate ratios estimated using Poisson regression ranged from 2.97 to 11.32 across sites, demonstrating the robustness of the signal for an increased risk (i.e., >1) across all sites. We did not conduct formal testing or sequential monitoring within each site, and thus only rate ratio effect estimates and standard errors are presented by site.

Table 17. Site-specific results of adjusted sequential monitoring using the Regression method comparing ACE inhibitors (ACEIs) to Beta Blockers (BBs) on angioedema risk over first 30 days of use, in Analysis 1

Site	BB N	BB Event (Rate)	ACEI N	ACEI Event(Rate)	BB Adj Rate	ACEI Adj Rate	Adj RR(SE)
1	20334	1(0.8)	26070	14(7.9)	0.7	8.2	11.32(2.85)
3	22391	4(2.8)	29923	21(10.4)	2.6	10.9	4.16(1.75)
15	5102	3(8.8)	7509	12(23.4)	8.3	24.8	2.97(1.94)

Site-specific adjusted Poisson regression model (no Sequential).
Covariates included sex, age, prescription NSAIDS use, oral corticosteroid use, allergic reactions, diabetes, ischemic heart disease, and any inpatient hospital stay
Abbreviations: Event(Rate)=Number(Rate per 1000 person-years) of angioedema within look and covariate category, Adj Rate=Adjusted Rate from site-specific adjusted Poisson regression model assuming entire site population was either exposed or unexposed, Adj RR(SE)=Adjusted Rate Ratio (Standard Error) comparing ACEI to BB from site-specific Poisson regression model.

Analysis 2 – Regression analysis with 360-day follow-up

Although the risk of angioedema is highest within the first 30 days of ACE inhibitor exposure, it can also occur after 30 days. To quantify this potentially longer-term increase in risk, we conducted a Regression analysis of angioedema defined to occur within a 360-day risk interval after ACE inhibitor initiation. As in the 30-day Regression analysis, the 360-day Regression analysis signaled at the first analysis time point. Thus, all results summarized in this section are based on this first (and only) sequential look.

Patient Characteristics

As shown in **Table 18**, the size and characteristics of the Analysis 2 cohort are exactly the same as the data shown in **Table 15** for the Analysis 1 cohort. There were 111,329 users of ACE inhibitors and BBs at the time of the first (and only) look.

Table 18: Baseline demographics of surveillance cohort included in Analysis 2, by ACE inhibitor (ACEI) and Beta Blocker (BB) exposure at the final look, total sample size=111,329

	Total (n= 111,329) N (%)	BB (n= 47,827) N (%)	ACEI (n= 63,502) N (%)
Sex			
Male	56075 (50.4)	20598 (43.1)	35477 (55.9)
Female	55254 (49.6)	27229 (56.9)	28025 (44.1)
Age			
18-44	35933 (32.3)	17339 (36.3)	18594 (29.3)
45-54	33013 (29.7)	12712 (26.6)	20301 (32.0)
55-64	26975 (24.2)	10520 (22.0)	16455 (25.9)
65-99	15408 (13.8)	7256 (15.2)	8152 (12.8)

	Total (n= 111,329) N (%)	BB (n= 47,827) N (%)	ACEI (n= 63,502) N (%)
Prescription NSAIDS			
No	97464 (87.5)	41743 (87.3)	55721 (87.7)
Yes	13865 (12.5)	6084 (12.7)	7781 (12.3)
Oral Corticosteroids			
No	103184 (92.7)	43757 (91.5)	59427 (93.6)
Yes	8145 (7.3)	4070 (8.5)	4075 (6.4)
Allergic Reactions			
No	99418 (89.3)	42189 (88.2)	57229 (90.1)
Yes	11911 (10.7)	5638 (11.8)	6273 (9.9)
Diabetes			
No	93115 (83.6)	43362 (90.7)	49753 (78.3)
Yes	18214 (16.4)	4465 (9.3)	13749 (21.7)
Heart Failure			
No	108635 (97.6)	46104 (96.4)	62531 (98.5)
Yes	2694 (2.4)	1723 (3.6)	971 (1.5)
Ischemic Heart Disease			
No	102021 (91.6)	41577 (86.9)	60444 (95.2)
Yes	9308 (8.4)	6250 (13.1)	3058 (4.8)
1+ Inpatient Hospital Stay			
No	99154 (89.1)	39787 (83.2)	59367 (93.5)
Yes	12175 (10.9)	8040 (16.8)	4135 (6.5)
Site			
1	46404 (41.7)	20334 (42.5)	26070 (41.1)
3	52314 (47.0)	22391 (46.8)	29923 (47.1)
15	12611 (11.3)	5102 (10.7)	7509 (11.8)

Sequential Results from the Regression Method

The 360-day Regression analysis signaled at the first of four planned analyses (**Table 19**). As anticipated, at the time of the identified signal there were more angioedema events observed in the 360-day Analysis 2 compared to the 30-day Analysis 1 (i.e., 134 total events for Analysis 2 and 55 events for Analysis 1, respectively). This is due to the longer follow-up time for Analysis 2. Specifically, 107 angioedema events were observed among 64,178 ACE inhibitor users compared with 27 angioedema events among 48,358 BB users at the time of the signal at look 1. The adjusted rate of angioedema per 1,000 person-years among ACE inhibitors users was estimated to be 2.73 times higher than for BB users (4.99 versus 1.83 per 1,000 person-years, respectively) using the Regression method.

Table 19: Results of adjusted sequential monitoring using the regression method comparing ACE inhibitors (ACEIs) to Beta Blockers (BBs) on angioedema risk over first 360 days of use, in Analysis 2

Look	Days	BB N	BB Event (Rate)	ACEI N	ACEI Event (Rate)	BB Adj Rate	ACEI Adj Rate	Adj RR	Score Test	Boundary	Error Spent	Signal
1	450	47827	27(1.96)	63502	107(4.78)	1.85	5.04	2.73	5.12	2.88	0.000	Yes

Adjusted Poisson regression model applied using GEE framework with sequential monitoring boundaries based on permutations. Covariates included sex, age, prescription NSAIDS use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic

heart disease, any inpatient hospital stay, and site.
Abbreviations: Event(Rate)=Number(Rate per 1000 person-years) of angioedema within look and covariate category, Adj Rate=Adjusted risk (per 1000 person-years) from adjusted Poisson regression model assuming entire population was either exposed or unexposed, Adj RR=Adjusted Rate Ratio comparing ACEI to BB from Poisson regression model, Score Test=Score Test statistic from GEE Poisson regression model, and Boundary=Sequential Boundary to compare the Score test estimate.

Note in column 2 of Table 19 that the timing of the first look for the Analysis 2 cohort was shifted (to 450 days) compared to Analyses 1 (at 90 days). This was done because the primary scientific question of interest in Analysis 2 was to compare the risk of angioedema *after 360 days* of exposure use. To address this question, we must first wait for the first quarter's worth of new users (i.e., those users who accrued into the surveillance cohort in the first 90 days) to have the opportunity for 360 days of follow-up time to be observed. Subjects may contribute less than 360 days due to censoring. Thus, instead of conducting the first look for this cohort on day 90 (when cohort members would have had only had a maximum of 90 days of exposure use), the first look was conducted *360 days after* day 90 (i.e., on day 450). Had there been no signal at this first look, the remaining planned analyses would have been conducted quarterly thereafter (i.e., on days 540, 630, and 720). If we had not shifted the timing of looks by 360 days and simply analyzed the cohort on a quarterly basis beginning immediately after the surveillance start date as in the 30-day in Analysis 1 (i.e., on days 90, 180, 270, and 360), then the estimated rate ratio at each look would *not* reflect average differences in angioedema risk at 360 days. It would instead be heavily weighted to reflect differences in risk for much shorter durations of time since most patients would only have had the opportunity to be observed for a much shorter duration than 360 days.

Site-Specific Results

As in Analysis 1, we examined results within each Data Partner subgroup at the time of the observed signal (**Table 20**). Adjusted event rates among ACE inhibitor users (4.15 5.16, and 7.36 per 1,000 person-years) and BB users (1.94, 1.71, and 2.15 per 1,000 person-years) varied somewhat across sites but were consistently higher among ACE inhibitor users within each site. The site-specific adjusted rate ratios estimated using Poisson regression ranged from 2.14 to 3.43 across sites, demonstrating the robustness of evidence for an increased risk (i.e., >1) across all sites.

Table 20. Site-specific results of adjusted sequential monitoring using the Regression method comparing ACE inhibitors (ACEIs) to Beta Blockers (BBs) on angioedema risk over first 360 days of use, in Analysis 2

Site	BB N	BB Event (Rate)	ACEI N	ACEI Event(Rate)	BB Adj Rate	ACEI Adj Rate	Adj RR(SE)
1	20334	11(1.94)	26070	36(4.1)	1.96	4.20	2.14(1.43)
3	22391	11(1.86)	29923	45(4.5)	1.72	5.22	3.03(1.42)
15	5102	5(2.29)	7509	26(7.04)	2.17	7.45	3.43(1.65)

Site-specific adjusted Poisson regression model (no Sequential).
Covariates included sex, age, prescription NSAIDS use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, and any inpatient hospital stay
Abbreviations: Event(Rate)=Number(Rate per 1000 person-years) of angioedema within look and covariate category, Adj Rate=Adjusted Rate from site-specific adjusted Poisson regression model assuming entire site population was either exposed or unexposed, Adj RR(SE)=Adjusted Rate Ratio (Standard Error) comparing ACEI to BB from site-specific Poisson regression model.

Analysis 3 – IPTW analysis with 30-day follow-up and a minimum of 30-day exposure duration requirement

Patient Characteristics

As detailed in the statistical analysis plan, we also conducted an analysis to estimate the difference in risk of angioedema between ACE inhibitor and BB users. Since the data used for the IPTW method was not returned by Data Partner 3 due to an error, the Analysis 3 cohort included Site 1, 4,and 15. The Analysis 3 cohort included 43,636 total new users; 59% of which were ACE inhibitor users (**Table 21**). As in Analysis 1, ACE inhibitor users were more likely than BB users to be male, 45 years of age or older, and have a diagnosis of diabetes. ACE inhibitor users were less likely than BB users to have a prior diagnosis of ischemic heart disease and an inpatient stay in the prior 12 months.

Table 21: Baseline demographics of surveillance cohort included in Analysis 3, by ACE inhibitor (ACEI) and Beta Blocker (BB) exposure at the final look at the final look, total sample size=43,636

	Total	BB	ACEI
Sex			
Male	21805 (50.0)	7620 (42.8)	14185 (54.9)
Female	21831 (50.0)	10179 (57.2)	11652 (45.1)
Age			
18-44	12470 (28.6)	5827 (32.7)	6643 (25.7)
45-54	12229 (28.0)	4438 (24.9)	7791 (30.2)
55-64	10814 (24.8)	3877 (21.8)	6937 (26.8)
65-99	8123 (18.6)	3657 (20.5)	4466 (17.3)
NSAIDS			
No	37920 (86.9)	15372 (86.4)	22548 (87.3)
Yes	5716 (13.1)	2427 (13.6)	3289 (12.7)
Aspirin			
No	43261 (99.1)	17610 (98.9)	25651 (99.3)
Yes	375 (0.9)	189 (1.1)	186 (0.7)
COX2 Inhibitors			
No	43193 (99.0)	17598 (98.9)	25595 (99.1)
Yes	443 (1.0)	201 (1.1)	242 (0.9)
Oral Corticosteroids			
No	40744 (93.4)	16377 (92.0)	24367 (94.3)
Yes	2892 (6.6)	1422 (8.0)	1470 (5.7)
Allergic Reactions			
No	39266 (90.0)	15748 (88.5)	23518 (91.0)
Yes	4370 (10.0)	2051 (11.5)	2319 (9.0)
Diabetes			
No	36676 (84.0)	16216 (91.1)	20460 (79.2)
Yes	6960 (16.0)	1583 (8.9)	5377 (20.8)
Heart Failure			
No	42557 (97.5)	17108 (96.1)	25449 (98.5)
Yes	1079 (2.5)	691 (3.9)	388 (1.5)
Ischemic Heart Disease			
No	40328 (92.4)	15579 (87.5)	24749 (95.8)
Yes	3308 (7.6)	2220 (12.5)	1088 (4.2)
1+ Inpatient Hospital Stay			

	Total	BB	ACEI
No	39040 (89.5)	14776 (83.0)	24264 (93.9)
Yes	4596 (10.5)	3023 (17.0)	1573 (6.1)
Site			
1	29402 (67.4)	12322 (69.2)	17080 (66.1)
4	2318 (5.3)	781 (4.4)	1537 (5.9)
15	11916 (27.3)	4696 (26.4)	7220 (27.9)

Sequential Results from the IPTW Method

A total of 30 angioedema events were observed; 26 were in the ACE inhibitor user group yielding adjusted risk estimates of 0.094% and 0.019% for ACE inhibitor and BB users, respectively (**Table 22**). The adjusted risk difference was estimated to be 0.075% and was statistically significant, generating a signal at look 1.

Table 22: Results of adjusted sequential monitoring using the IPTW method comparing ACE inhibitors (ACEIs) to Beta Blockers (BBs) on angioedema risk over first 30 days of use, in Analysis 3

Look	Days	BB N	BB Event (%)	ACEI N	ACEI Event (%)	BB Adj %	ACEI Adj %	Adj RD	IPTW Test	Boundary	Error Spent	Signal
1	90	17799	4 (0.022)	25837	26 (0.101)	0.019	0.094	0.075	3.517	3.378	0.000	Yes

Adjusted stratified risk difference model applied using IPTW with sequential monitoring boundaries based on permutations.
Covariates included sex, age, prescription NSAIDS use, aspirin use, COX2 inhibitor use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Event (%)=Number(Risk %) of angioedema within look and covariate category, Adj %=Adjusted Risk % from stratified IPTW model for a given exposure group, Adj RD=ACEI Adj % – BB Adj %, IPTW Test=Adj RD/Standard Error(Adj RD), and Boundary=Sequential Boundary to compare the IPTW test estimate.

Site-Specific Results

We performed site-specific adjusted weighted linear regression to explore whether the association between ACE inhibitors and angioedema risk varied across sites. The adjusted risk difference for all 3 sites were greater than 0 indicating ACE inhibitor users at all three sites had a higher risk of angioedema comparing to BB users (**Table 23**). The magnitude of the risk difference varied from 0.06 to 0.26% across sites, which is not surprising given the very few number of events that were observed within each site.

Table 23. Site-specific results of adjusted sequential monitoring using the IPTW method comparing ACE inhibitors (ACEIs) and Beta Blockers (BBs) on angioedema risk over first 30 days of use, in Analysis 3

Site	BB N	BB Event (%)	ACEI N	ACEI Event(%)	BB Adj %Event	ACEI Adj %Event	Adj RD
1	12322	1 (0.008)	17080	12 (0.070)	0.006	0.067	0.06
4	781	0 (0.00)	1537	4 (0.260)	0.000	0.259	0.26
15	4696	3 (0.064)	7220	10 (0.139)	0.055	0.129	0.08

Adjusted risk difference model applied using IPTW for each site (no Sequential).
Covariates included sex, age, prescription NSAIDS use, aspirin use, COX2 inhibitor use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Event(%)=Number(Risk %) of angioedema within look and covariate category, Adj %Event= Adjusted Risk % from site-specific IPTW model for a given exposure group, Adj RD= site-specific ACEI Adj %Event - BB Adj %Event

Analysis 4 –IPTW analysis with 30-day follow-up and no minimum exposure duration requirement

Patient Characteristics

The Analysis 4 cohort was slightly larger than the Analysis 3 cohort since there was no minimum exposure duration requirement in the Analysis 4 cohort (vs 30 days requirement in the Analysis 3 cohort). A total of 47,073 individuals were included, and 58% were ACE inhibitor users (**Table 24**). Patient characteristics were similar to those in the Analysis 3 cohort.

Table 24: Baseline demographics of surveillance cohort included in Analysis 4, by ACE inhibitor (ACEI) and Beta Blocker (BB) exposure at the final look, total sample size=47,073

	Total (n= 47073) N (%)	BB (n= 19911) N (%)	ACEI (n= 27162) N (%)
Sex			
Male	23506 (49.9)	8607 (43.2)	14899 (54.9)
Female	23567 (50.1)	11304 (56.8)	12263 (45.1)
Age			
18-44	13636 (29.0)	6667 (33.5)	6969 (25.7)
45-54	13138 (27.9)	4937 (24.8)	8201 (30.2)
55-64	11569 (24.6)	4284 (21.5)	7285 (26.8)
65-99	8730 (18.5)	4023 (20.2)	4707 (17.3)
NSAIDS			
No	40923 (86.9)	17244 (86.6)	23679 (87.2)
Yes	6150 (13.1)	2667 (13.4)	3483 (12.8)
Aspirin			
No	46679 (99.2)	19712 (99.0)	26967 (99.3)
Yes	394 (0.8)	199 (1.0)	195 (0.7)
COX2 Inhibitors			
No	46606 (99.0)	19693 (98.9)	26913 (99.1)
Yes	467 (1.0)	218 (1.1)	249 (0.9)
Oral Corticosteroids			
No	43933 (93.3)	18336 (92.1)	25597 (94.2)
Yes	3140 (6.7)	1575 (7.9)	1565 (5.8)
Allergic Reactions			
No	42364 (90.0)	17626 (88.5)	24738 (91.1)
Yes	4709 (10.0)	2285 (11.5)	2424 (8.9)
Diabetes			
No	39659 (84.2)	18130 (91.1)	21529 (79.3)
Yes	7414 (15.8)	1781 (8.9)	5633 (20.7)
Heart Failure			
No	45827 (97.4)	19116 (96.0)	26711 (98.3)
Yes	1246 (2.6)	795 (4.0)	451 (1.7)
Ischemic Heart Disease			
No	43404 (92.2)	17446 (87.6)	25958 (95.6)
Yes	3669 (7.8)	2465 (12.4)	1204 (4.4)
1+ Inpatient Hospital Stay			
No	41944 (89.1)	16507 (82.9)	25437 (93.6)
Yes	5129 (10.9)	3404 (17.1)	1725 (6.4)

	Total (n= 47073) N (%)	BB (n= 19911) N (%)	ACEI (n= 27162) N (%)
Site			
1	31899 (67.8)	13918 (69.9)	17981 (66.2)
4	2513 (5.3)	882 (4.4)	1631 (6.0)
15	12661 (26.9)	5111 (25.7)	7550 (27.8)

Sequential Results from the IPTW Method

Sequential results from the IPTW method for Analysis 4 were similar to those in Analysis 3 when a minimum of 30 days exposure duration was required. As in Analysis 3, a signal with an adjusted risk difference of 0.07% was detected at the first planned look (**Table 25**).

Table 25: Results of adjusted sequential monitoring using the IPTW method comparing ACE inhibitors(ACEIs) to Beta Blockers (BBs) on angioedema risk over first 30 days of use, in Analysis 4

Look	Days	BB N	BB Event (%)	ACEI N	ACEI Event (%)	BB Adj %	ACEI Adj %	Adj RD	IPTW Test	Boundary	Error Spent	Signal
1	90	19911	4 (0.020)	27162	26 (0.096)	0.017	0.089	0.072	3.593	3.398	0.000	Yes

Adjusted stratified risk difference model applied using IPTW with sequential monitoring boundaries based on permutations.
Covariates included sex, age, prescription NSAIDS use, aspirin use, COX2 inhibitor use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Event (%)=Number(Risk %) of angioedema within look and covariate category, Adj %=Adjusted Risk % from stratified IPTW model for a given exposure group, Adj RD=ACEI Adj % – BB Adj %, IPTW Test=Adj RD/Standard Error(Adj RD), and Boundary=Sequential Boundary to compare the IPTW test estimate.

Site-Specific Results

Point estimates for the site-specific analyses were also very similar to those in Analysis 3. The adjusted risk difference for all 3 sites were greater than 0 indicating that ACE inhibitor users at all three sites had a higher risk of angioedema compared to BB users (**Table 26**).

Table 26. Site-specific results of adjusted sequential monitoring using the IPTW method comparing ACE inhibitors(ACEIs) to Beta Blockers (BBs) on angioedema risk over first 30 days of use, in Analysis 4

Site	BB N	BB Event (%)	ACEI N	ACEI Event(%)	BB Adj %Event	ACEI Adj %Event	Adj RD
1	13918	1(0.007)	17981	12(0.067)	0.006	0.064	0.06
4	882	0(0.000)	1631	4(0.245)	0	0.243	0.24
15	5111	3(0.059)	7550	10(0.132)	0.05	0.123	0.07

Adjusted risk difference model applied using IPTW for each site (no Sequential).
Covariates included sex, age, NSAIDS use, aspirin use, COX2 inhibitor use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Event(%)=Number(Risk %) of angioedema within look and covariate category, Adj %Event= Adjusted Risk % from site-specific IPTW model for a given exposure group, Adj RD= site-specific ACEI Adj %Event - BB Adj %Event

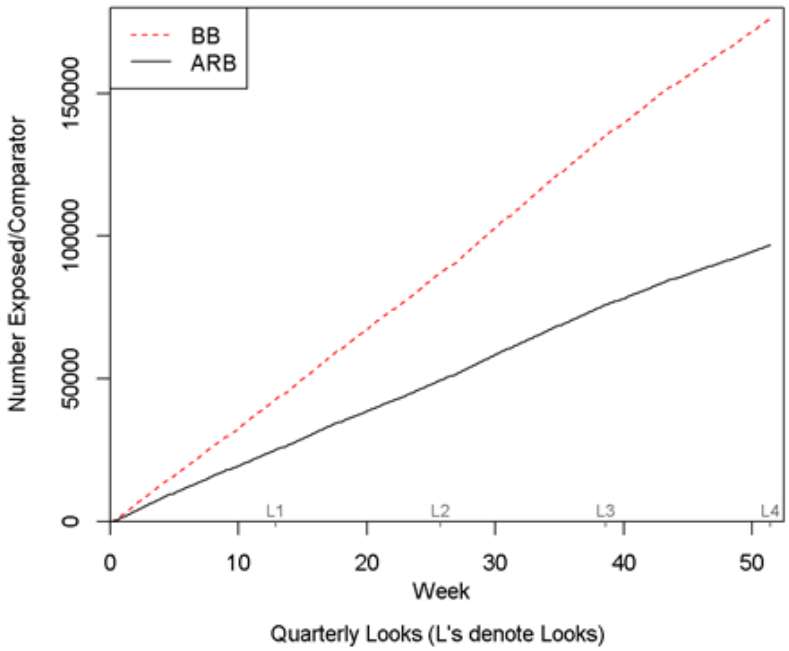
2. ARBs and risk of angioedema

Since all four analyses for ACE inhibitor example signaled at the first look, we present results of selected analyses for a second example: ARBs and angioedema risk. Specifically, we conducted and report results from Analysis 1 (Regression analysis with a 30-day follow-up) and Analysis 3 (IPTW analysis with 30-day

follow-up and a minimum of 30-day exposure duration requirement) for ARBs and angioedema risk. For both of these ARBs analyses, no signal was detected at any of the 4 planned looks.

Uptake of BBs was sizeable and steady among the surveillance cohort, but ARB uptake was smaller and slower (**Figure 6**). For the Regression method, Data Partner sites 4 and 15 were dropped due to a lack of events in either the ARB or BB group. For the IPTW method, site 3 did not return data due to an error. Thus, the IPTW analysis excluded site 3 due to an error and site 4 due to a lack of events.

Figure 6. Total uptake of ARB and BB overtime at the final look



Analysis 1 - Regression analysis with 30-day follow-up

Patient Characteristics

In **Table 27**, we describe patient characteristics for the ARB and BB cohorts included in Analysis 1 that evaluated risk of angioedema in the first 30-days of exposure at the fourth and final look. At the final look, 273,381 ARB (35.4%) and BB (64.6%) users were included in the analysis. ARB users differed from BB users in that ARB users were more likely to be male, 45 years of age or older and have a diagnosis of diabetes. ARB users were less likely than BB users to have heart failure, ischemic heart disease, or an inpatient stay in the prior year (**Table 27**).

Table 27: Baseline demographics of surveillance cohort included in Analysis 1, by BB and ARB exposure groups at the final look, total sample size=273,381

	Total (n=273381) N (%)	BB (n=176511) N (%)	ARB (n=96870) N (%)
Sex			
Male	125186 (45.8)	76022 (43.1)	49164 (50.8)
Female	148195 (54.2)	100489 (56.9)	47706 (49.2)

	Total (n=273381) N (%)	BB (n=176511) N (%)	ARB (n=96870) N (%)
Age			
18-44	89042 (32.6)	64483 (36.5)	24559 (25.4)
45-54	80076 (29.3)	47812 (27.1)	32264 (33.3)
55-64	66400 (24.3)	39325 (22.3)	27075 (27.9)
65-99	37863 (13.8)	24891 (14.1)	12972 (13.4)
NSAIDS			
No	240747 (88.1)	155205 (87.9)	85542 (88.3)
Yes	32634 (11.9)	21306 (12.1)	11328 (11.7)
Oral Corticosteroids			
No	250628 (91.7)	161151 (91.3)	89477 (92.4)
Yes	22753 (8.3)	15360 (8.7)	7393 (7.6)
Allergic Reactions			
No	241310 (88.3)	155931 (88.3)	85379 (88.1)
Yes	32071 (11.7)	20580 (11.7)	11491 (11.9)
Diabetes			
No	238845 (87.4)	159583 (90.4)	79262 (81.8)
Yes	34536 (12.6)	16928 (9.6)	17608 (18.2)
Heart Failure			
No	265503 (97.1)	170193 (96.4)	95310 (98.4)
Yes	7878 (2.9)	6318 (3.6)	1560 (1.6)
Ischemic Heart Disease			
No	244514 (89.4)	153243 (86.8)	91271 (94.2)
Yes	28867 (10.6)	23268 (13.2)	5599 (5.8)
1+ Inpatient Hospital Stay			
No	240505 (88.0)	148456 (84.1)	92049 (95.0)
Yes	32876 (12.0)	28055 (15.9)	4821 (5.0)
Site			
1	135259 (49.5)	85909 (48.7)	49350 (50.9)
3	138122 (50.5)	90602 (51.3)	47520 (49.1)

Sequential Results from the Regression Method

A summary of the results at each of the four sequential looks is presented in **Table 28**. At the final analysis, 29 and 48 angioedema events were observed in the ARB and BB user groups, respectively. The adjusted angioedema risk within the first 30 days of exposure in new users of ARBs was 32% higher than the risk among users of BBs (i.e., adjusted RR=1.32). However, the test statistic did not cross the pre-specified sequential monitoring threshold, and thus no signal was detected. The adjusted estimates of the rate of angioedema in the ARB and BB groups were 4.42 and 3.36 events per 1000 person-years, respectively.

Table 28. Results of adjusted sequential monitoring using the Regression method comparing ARB to BB on angioedema risk over first 30 days of use, in Analysis 1

Look	Days	BB N	BB Event (Rate)	ARB N	ARB Event (Rate)	BB Adj Rate	ARB Adj Rate	Adj RR	Score Test	Boundary	Error Spent	Signal
1	90	42725	6(2.15)	25121	9(5.23)	2.21	5.00	2.27	1.53	3.70	0.000	No
2	180	87314	19(3.03)	49529	20(5.39)	2.78	6.43	2.32	2.45	2.61	0.003	No
3	270	135384	33(3.31)	76015	24(4.12)	3.07	4.77	1.55	1.56	1.64	0.046	No
4	360	176511	48(3.62)	96870	29(3.81)	3.36	4.42	1.32	1.14	2.02	0.048	No

Adjusted Poisson regression model applied using GEE framework with sequential monitoring boundaries based on permutations. Covariates included sex, age, prescription NSAIDS use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and site.
Abbreviations: Event(Rate)=Number(Rate per 1000 person-years) of angioedema within look and covariate category, Adj Rate=Adjusted risk (per 1000 person-years) from adjusted Poisson regression model assuming entire population was either exposed or unexposed, Adj RR=Adjusted Rate Ratio comparing ACEI to BB from Poisson regression model, Score Test=Score Test statistic from GEE Poisson regression model, and Boundary=Sequential Boundary to compare the Score test estimate.
Sequential p-value for signal: 0

Site-Specific Results

Results from the Regression method were similar between the two study sites for the ARB analysis (**Table 29**).

Table 29. Site-specific results of adjusted sequential monitoring using the Regression method comparing ARB and BB users on angioedema risk over first 30 days of use, in Analysis 1

Site	BB N	BB Event (Rate)	ARB N	ARB Event(Rate)	BB Adj Rate	ARB Adj Rate	Adj RR(SE)
1	85909	22(3.40)	49350	15(3.87)	3.19	4.35	1.36(1.41)
3	90602	26(3.83)	47520	14(3.76)	3.55	4.41	1.24(1.41)

Site-specific adjusted Poisson regression model (no Sequential).
Covariates included sex, age, NSAIDS use, oral corticosteroid use, allergic reactions, diabetes, ischemic heart disease, and any inpatient hospital stay
Abbreviations: Event(Rate)=Number(Rate per 1000 person-years) of angioedema within look and covariate category, Adj Rate=Adjusted Rate from site-specific adjusted Poisson regression model assuming entire site population was either exposed or unexposed, Adj RR(SE)=Adjusted Rate Ratio (Standard Error) comparing ARB to BB from site-specific Poisson regression model.

<u>Analysis 3 – IPTW analysis with 30-day follow-up and a minimum of 30-day exposure duration requirement</u>

Patient Characteristics

As detailed in the statistical analysis plan, we conducted an analysis to estimate the difference in risk of angioedema between users of ARBs and BBs. The Analysis 3 cohort included 134,082 total new users across the two sites; 34% were ARB users (**Table 30**). Similar to Analysis 1, ARB users differed from BB users in that ARB users were more likely to be male, 45 years of age or older and have a diagnosis of diabetes. ARB users were less likely than BB users to have heart failure, ischemic heart disease, or an inpatient stay in the prior year (**Table 30**).

Table 30: Baseline demographics of surveillance cohort included in Analysis 3, by BB and ARB exposure groups at the final look, total sample size=134082

	Total (n=134082) N (%)	BB (n=88990) N (%)	ARB (n=45092) N (%)
Sex			
Male	60672 (45.2)	38046 (42.8)	22626 (50.2)
Female	73410 (54.8)	50944 (57.2)	22466 (49.8)
Age			
18-44	39320 (29.3)	29069 (32.7)	10251 (22.7)
45-54	37194 (27.7)	22682 (25.5)	14512 (32.2)
55-64	32458 (24.2)	19629 (22.1)	12829 (28.5)
65-99	25110 (18.7)	17610 (19.8)	7500 (16.6)
NSAIDS			
No	117685 (87.8)	77676 (87.3)	40009 (88.7)
Yes	16397 (12.2)	11314 (12.7)	5083 (11.3)
Aspirin			
No	133291 (99.4)	88351 (99.3)	44940 (99.7)
Yes	791 (0.6)	639 (0.7)	152 (0.3)
COX2			
No	132452 (98.8)	87993 (98.9)	44459 (98.6)
Yes	1630 (1.2)	997 (1.1)	633 (1.4)
Oral Corticosteroids			
No	123955 (92.4)	81997 (92.1)	41958 (93.0)
Yes	10127 (7.6)	6993 (7.9)	3134 (7.0)
Allergic Reactions			
No	119484 (89.1)	79398 (89.2)	40086 (88.9)
Yes	14598 (10.9)	9592 (10.8)	5006 (11.1)
Diabetes			
No	117395 (87.6)	80829 (90.8)	36566 (81.1)
Yes	16687 (12.4)	8161 (9.2)	8526 (18.9)
Heart Failure			
No	129968 (96.9)	85656 (96.3)	44312 (98.3)
Yes	4114 (3.1)	3334 (3.7)	780 (1.7)
Ischemic Heart Disease			
No	120390 (89.8)	77908 (87.5)	42482 (94.2)
Yes	13692 (10.2)	11082 (12.5)	2610 (5.8)
1+ Inpatient Hospital Stay			
No	117425 (87.6)	74569 (83.8)	42856 (95.0)
Yes	16657 (12.4)	14421 (16.2)	2236 (5.0)
Site			
1	113861 (84.9)	70314 (79.0)	43547 (96.6)
15	20221 (15.1)	18676 (21.0)	1545 (3.4)

Sequential Results from the IPTW method

A summary of the IPTW results at each of the four sequential looks is presented in **Table 31**. At the final analysis (Look 4), 15 and 39 angioedema cases were observed in the ARB and BB user groups, respectively. The adjusted risk was 0.034% and 0.038% for ARB and BB users, respectively. This yielded an adjusted risk difference of -0.003 which was not statistically significant and thus did not generate a signal.

Table 31: Results of adjusted sequential monitoring using the IPTW method comparing ARB to BB on angioedema risk over first 30 days of use, in Analysis 3

Look	Days	BB N	BB Event (%)	ARB N	ARB Event (%)	BB Adj %	ARB Adj %	Adj RD	IPTW Test	Boundary	Error Spent	Signal
1	90	17018	4 (0.024)	8581	4 (0.047)	0.018	0.068	0.050	1.119	3.449	0.000	No
2	180	40376	13 (0.032)	20421	9 (0.044)	0.025	0.050	0.024	1.071	2.240	0.000	No
3	270	65087	25 (0.038)	33454	12 (0.036)	0.032	0.040	0.008	0.495	1.834	0.004	No
4	360	88990	39 (0.044)	45092	15 (0.033)	0.038	0.034	-0.003	-0.257	1.340	0.050	No

Adjusted stratified risk difference model applied using IPTW with sequential monitoring boundaries based on permutations.
Covariates included sex, age, NSAIDS use, aspirin use, COX2 inhibitor use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Event(%)=Number(Risk %) of angioedema within look and covariate category, Adj %=Adjusted Risk % from stratified IPTW model for a given exposure group, Adj RD=ARB Adj % – BB Adj %, IPTW Test=Adj RD/Standard Error(Adj RD), and Boundary=Sequential Boundary to compare the IPTW test estimate.

Site-Specific Results

Results from the IPTW method differed between the two study sites for the ARB analysis. Site 1 showed an adjusted risk estimate greater than zero, indicating ARB users had a higher risk of angioedema comparing to BB users (**Table 32**). Results from both sites but especially site 15 should be interpreted with caution given the small number of events in the ARB users.

Table 32. Site-specific results of adjusted sequential monitoring using the IPTW method comparing ARB and BB users on angioedema risk over first 30 days of use, in Analysis 3

Site	BB N	BB Event (%)	ARB N	ARB Event(%)	BB Adj %Event	ARB Adj %Event	Adj RD
1	70314	22 (0.031)	43547	14 (0.032)	0.028	0.030	0.003
15	18676	17 (0.091)	1545	1 (0.065)	0.094	0.056	-0.037

Adjusted risk difference model applied using IPTW for each site (no Sequential).
Covariates included sex, age, NSAIDS use, aspirin use, COX2 inhibitor use, oral corticosteroid use, allergic reactions, diabetes, heart failure, ischemic heart disease, any inpatient hospital stay, and indicator for each look within site strata.
Abbreviations: IPTW=Inverse Probability of Treatment Weighting, Event(%)=Number(Risk %) of angioedema within look and covariate category, Adj %Event= Adjusted Risk % from site-specific IPTW model for a given exposure group, Adj RD= site-specific ARB Adj %Event - BB Adj %Event

D. SUMMARY

1. ACE Inhibitors

Our surveillance plan on the risk of angioedema with ACE inhibitors vs. BBs was successful in signaling at the first look for all 4 methods. Our results are validated by a much larger study in Sentinel where ACE

inhibitors were associated with an increase in the risk of angioedema.[56] In a sample size of 1.8 million ACE inhibitor users and 1.6 million BB users with a maximum follow-up of 365 days, Toh et al. reported an overall adjusted hazard ratio that indicated approximately 3 times (HR=3.04; 95% CI, 2.81-3.27) the rate of angioedema in ACE inhibitor users compared to users of BBs.[53] Toh et al. reported an adjusted 30 day rate of 9.7 per 1000 PY for ACE inhibitor users, 2.98 per 1000 PY for BB users, and a hazard ratio of 3.57 for ACE inhibitors compared to BBs.[53] Toh et. al. used a case-centered approach with propensity scores that included the entire cohort. Our estimated risk ratio at 360 days for ACE inhibitors vs. BB (RR=2.73) is similar to that of Toh but our 30 day risk (RR=4.6) is higher than Toh's estimate. Given that our estimate occurs at the time of a signal, when only a fraction of the total sample size has accrued, there is potential for some instability and higher variability in the estimated effect. Over time we would expect the effect estimate to stabilize to a level closer to that presented in Toh, et. al . Our rate of angioedema at 30 days in the ACE inhibitor group was close (11.6 per 1000 PY) to Toh's estimate but not the same. However, our rate of angioedema at 30 days in the BB group (2.46 per 1000 PY) was very similar to Toh's estimate. The variation in risk across sites in the 30 day regression analysis (Analysis 1) is a limitation but there was little variation by site in the 360 day analysis (Analysis 2) and observed variations in the 30 day analysis are likely due to the small number of events.

Toh et. al. did not report risk differences but we estimated the unadjusted risk difference from the reported cumulative incidence rates and the adjusted risk difference from the propensity score adjusted hazard ratio and cumulative incidence for BB users .[53] As such, we estimate the unadjusted risk difference in the Toh analyses is 0.054% and the adjusted risk difference is 0.059% at 30 days. This is close to our 30 day unadjusted risk difference from analysis 3 (0.079%) and 4 (0.076%) and adjusted risk difference from analysis 3 (0.075%) and 4 (0.072%).

Characteristics of ACE inhibitor and BB users in our cohort are consistent with preferential use of ACE inhibitors in diabetic patients and BBs among patients with chronic heart disease and recent myocardial infarctions. The distributions are similar to the propensity score matched cohort described by Toh et al.[53] Our observation that the prevalence of diabetes in the ACE inhibitor group was double that of the BB group was also reported in Toh, et al, 19%, and 7%, respectively. However, our medication cohorts tended to be younger than those analyzed in Toh et al. [53] The number of males in our ACE inhibitor and BB groups were similar to that reported by Toh et. al (43% and 53%). The study by Toh and colleagues included a much larger sample size which may also account for observed differences.

The similarity between our findings across the various analyses/methods and compared to that of Toh and colleagues lend to confidence in the appropriateness of our methods.

2. ARBs

The analysis of ARBs versus BBs did not signal for any of the four looks, yielding an estimated relative risk of 1.32 at the final or fourth look (Analysis 1). This may be due to a variety of factors such as no or weak association and low rates of angioedema in users of ARBs. However, the study was powered on what was considered a meaningful difference in risk (RR=2.0). Our 30 day relative risk estimate is close to Toh, et al, where the 30-day rates of angioedema in the ARB and BB groups were 3.5 and 3.0 per 1000 person-years, respectively, with a data-partner adjusted hazard ratio of 1.4 and a propensity-score adjusted hazard ratio of 1.5.

Using a similar method as described above to calculate risk differences in the Toh et. al study, we estimated an unadjusted risk difference of 0.004% and an adjusted risk difference of 0.01% at 30 days. These are in the opposite direction yet relatively close to our estimates (-0.011% unadjusted and -

0.033% adjusted) in analysis 3. The small number of events may have limited the interpretation of our results for ARBs and our ability to compare to the Toh study. However as noted below, IPTW provides more stable estimates than GEE when events are rare.

As with the ACE inhibitor versus BB analysis, the patient characteristics of ARB and BB users in our cohort are consistent with preferential use of ARBs in diabetic patients and BBs among patients with chronic heart disease and recent myocardial infarctions, and are similar to the cohort described by Toh et al.[53] The number of males in our ARB and BB groups was similar to that reported by Toh et. al (43%, and 49%) but our cohort was younger than the population evaluated by Toh and colleagues. Our observation that the prevalence of diabetes in the ARB group was double that of the BB group was also reported in Toh, et al. at 16%, and 7%, respectively. Differences between our population and the population used in the study by Toh and colleagues are likely due to different data partners and time periods.

3. Difference in methods

Because the IPTW and regression analyses were not run on the same sites, we cannot directly compare results. We do expect their performance to differ in general, however. Signal detection is expected to be faster for IPTW, which uses a risk difference measure, compared with regression, which uses a relative measure. Signals can be detected more quickly using the risk difference because it is more stable than a ratio measure when events are rare. In particular, as the probability of an adverse event occurring decreases, the denominator of a ratio measure will be very small. This causes the ratio to be increasingly large and variable. The risk difference does not have this limitation. It is less variable and thus more powerful than a ratio measure when there are few events. This power differential between the IPTW risk difference method and the regression approach can be seen in the tables and figures presented in Section IV. These planning data showed that considerably smaller sample sizes are needed to signal with the risk difference IPTW method. In the ACE inhibitor and angioedema example, we based our testing plan on what is expected to be feasible within the Sentinel setting based on data updating by Data Partners (i.e., we implemented quarterly testing). If we had conducted testing more frequently and with smaller sample size increments at the first initial sequential looks, we would likely have seen an earlier signal for the IPTW risk difference approach compared to regression (or with any other method that use a ratio measure to quantity effects such as the hazard ratio estimated using a PS matched design). However, the IPTW method is best for exposure-outcome pairs where the time to event is expected to be relatively short.

4. Challenges

Results are limited by the sample size given the rarity of the outcome under study and excluding some sites from analyses due to no events in the therapeutic classes of interest or errors in the delivery of data. Our data source was administrative health plan records which are collected for administrative and not research purposes. Therefore, misclassification of outcomes, exposures, and confounders is possible. Cases of angioedema may have been missed under scenarios such as miscoding or cases not presenting for care (i.e., mild/quick resolving). We used dispensed prescriptions as a surrogate for medication exposure. Although this is one of the best available measures, it nevertheless has limitations. Patients may obtain drugs from non-pharmacy sources, such as samples provided by physicians, and medications may be started during a hospital stay, a period of medication use not ascertained in our data. Patients will not always adhere to the prescription regimen, leading to misclassification. While we adjusted for known confounders available in the Common Data Model,

residual confounding cannot be ruled out. For example, we lacked information on potential confounders such as race and smoking status (since identifying smoking status using ICD-9 codes suffers from misclassification). We also limited the number of confounders included in the GEE models due to potential issues with small cell sizes that can result when a large number of confounders are included in standard adjustment.

We created an artificial sequential data environment with one data pull to test two methods on example drug-outcome pairs. In real world sequential surveillance, data are continually updated and can change after planned refreshes of data. Further, our 360 day analysis is limited by an assumed constant risk of angioedema whereas the risk rate will be unknown for most drugs.

Since this was a methods demonstration project that re-analyzed a previously known safety association, we did not have the opportunity to examine the impact of dynamic data changes and quarterly data refreshes on the Regression or IPTW method in practice. Dynamic data updating and refreshing introduces several complications, which must be dealt with by any surveillance method that might be used. First, dynamic data updating can result in changes over time to the outcome, exposure, or baseline confounder status of individuals included in the cohort. As with any method, if the data change dramatically over time then the results from the Regression or IPTW method could change substantially. These changes to the results would appropriately reflect the substantive updates to the data. This problem can be minimized for any method, including Regression or IPTW, if a data lag is applied that allows the database information to 'settle' (i.e., allows the data to become more complete and stable before including data in an analysis). However, unlike other methods such as PS matching, small changes to the data over time would NOT be expected to substantively change the Regression and IPTW results. For instance, if exposure or confounder data change slightly, PSs can change slightly. With PS matching, even slight changes to the PS have been shown to alter a high proportion of the matches, which changes the composition of the comparators included in the matched cohort and, in turn, changes the matched pairs/sets and outcomes that are ultimately informative to the final effect estimate. With very few outcomes, such changes can produce substantive changes to effect estimates due to small sample variability. For Regression and IPTW, however, small changes to exposure or confounder data do not alter who is included in the cohort (since all comparators are included) or what subjects and outcomes are informative (since all subjects and outcomes inform the estimate). The modestly updated data are simply folded into the regression models and should yield accordingly modest updates to the effect estimates.

A second challenge is that data refreshing can lead to an inability to link a specific individual's data before and after the data refresh at some Sentinel Data Partners (e.g., it may not be possible to match the drug exposure time for a given individual before a data refresh to that same person's continued exposure time after the data refresh). This complication does NOT affect implementation of either the Regression or IPTW method. The Regression method does not use individual data but rather inputs aggregated information by exposure and confounder strata. Thus, it is not necessary to connect a particular individual's exposure time before and after a data refresh. The IPTW method does use individual data, but at each new analysis it does not need to 'look back' to the individual information prior to the last data refresh. Instead, at each new analysis, individual data at each Data Partner can be cumulatively refreshed from the surveillance start and then used to estimate the PS and risk difference within that Data Partner.

A final complication is that data refreshing, which occurs with variable timing across Data Partners based on administrative constraints unique to each Data Partner, can make it difficult to implement a sequential design testing schedule according to initial plan. This is because periodic data refreshing

results in an unpredictable amount of new information being available at each new data refresh. For example, if plans are made to analyze the data after we observe each new 10,000 users but a data refresh results in 15,000 new users being observed since the last analysis, the current analysis will include 5,000 more users than planned. Changes to the planned timing of sequential tests impacts the Type 1 error, which the surveillance team wants to control. Unexpected changes to the composition of the surveillance cohort (e.g., with respect to confounders) also impacts the Type 1 error. Both the Regression and IPTW method's signaling threshold computation approaches are designed to account for these unpredictable data issues. They do so by using the actual information about the observed cohort at a given analysis time point (i.e., the actual number of available cohort members and the actual confounder composition of the cohort at a given analysis instead of the expected/planned information at surveillance outset) to derive the signaling threshold. Doing this adjustment to the signaling threshold over time thereby correctly controls the Type 1 error based on the data that are actually observed.

Although we have described the expected performance of the Regression and IPTW methods in the context of dynamic data updating challenges, an important area for future work is to actually implement the Regression and IPTW in practice in a truly active surveillance setting where data are updated dynamically over time to confirm that method performs as anticipated.

5. Strengths

There are numerous theoretical advantages of using the IPTW approach and they briefly include (see Table 1 for more detail on both strengths and limitations):

- Confounding by site is accounted for via stratification, which effectively accounts for potential interactions between data partners and confounders.
- Use of a propensity score enables adjustment for a large number of confounders at each data partner.
- It accounts for differences in the variability of the estimated propensity score across sites (i.e., it reflects larger amounts of uncertainty in PS estimated at small sites versus larger ones).
- Risk differences are often conceptually appealing to policy-makers since they can be readily translated into the number of excess events due to the new product, making it simpler to weigh the risks and benefits. However, relative risks can also be approximated for comparability to other studies if needed.
- Updating confounders with planned refreshes is easy to operationalize in comparison with methods that match on confounders.

Strengths of standard regression include the following (see Table 1 for more detail on both strengths and limitations)

- Uses all event information from the cohort
- Applicable for chronic medication use and long-term outcomes which may account for a large portion of drug-outcome evaluations
- Allows for either standard adjustment of variables or PS adjustment (i.e., strength when there are a large number of confounders)
- Estimates a relative risk or odds ratio which are well understood by researchers and comparable to much of the literature
- Ease of conducting subgroup analyses

6. Enhancements

During data collection and analyses, enhancements were made to the existing code and reports for the Regression and IPTW methods by this workgroup. They include the following:

- Developed new within-Data-Partner SAS data aggregation macro for GSGEE with variable follow up time.
  - De-identifies data through standard aggregation by covariates.
  - Aggregates follow up time, number of events, and number of censorings in discrete time bins corresponding to the planned sequential analysis times.
  - For the ACE inhibitors vs beta blocker comparison, the individual level data across all 4 Data Partners contained 1.9 million records – the aggregated data contained approximately 130,000 records, or about 7% of the original.
- Developed new SAS macro for GSGEE to expand data aggregated with the use of the aggregation macro described above.
  - Creates individual level dataset from de-identified, aggregate data, which is important for correctly estimating robust standard errors which are used for inference with GSGEE.
  - Follow up times are averaged over "like" individuals in the same calendar time bins which correspond to the analysis times in the sequential surveillance plan.
  - Study start day for an individual is randomly generated within a known start window that is encoded as a variable in the aggregated data.
- Developed, tested and implemented new GSGEE analytic code to allow for the group sequential analysis of relative rates in studies where the amount of exposure varies across individuals, e.g., a new user cohort study where exposed individuals are followed until they have an outcome of interest, discontinue the exposure, are lost to follow up or are administratively censored at the end of some predefined study period.
- Updated SAS macros for IPTW to collect only summary information and aggregate data from each Data Partner, eliminating computationally intensive calculations for the sequential monitoring boundary. Boundary calculations are now done at the time of analysis instead of during the Data Partners' execution of analytic code.

7. Conclusion

Existing methods used for sequential design planning for randomized trials and observational safety surveillance assessments within the VSD and Mini-Sentinel provide a strong foundation upon which to build a more formal framework to plan future routine safety evaluations using electronic health care databases. In this section, we have provided recommendations on how practices from randomized trials can be adapted to accommodate the unique challenges of conducting safety surveillance activities in the observational setting of electronic health record databases, which contributes to an emerging literature on this topic.[57,58] We have also illustrated ways in which existing methods from observational settings like the VSD could be improved, by further leveraging well-established best practices from trial settings and tailoring them to meet the challenges posed by an electronic data environment.

This work points to three important sequential design steps that should be addressed during the planning phase for safety database surveillance activities. The first two points should be considered in both sequential and one-time surveillance activities. First, pre-specification of the surveillance design and analytic plan is not always possible but preferred. Second, use of existing data to inform surveillance

planning can reduce the number of assumptions that need to be made at the planning phase and, in turn, minimize downstream changes to initial sequential plans. Third, selection of a sequential design should include statistical evaluation and clear communication of the sequential design and analysis with all relevant stakeholders so that the operating characteristics are well understood in advance of implementation. In addition, due to the dynamic nature of the health care data sources, it is important that selected methods offer the ability to be flexible in their implementation and that investigators document any resulting changes to initial plans that are caused by unpredictable data. Ultimately, we hope that this work can spark further dialogue that will lead to more systematic sequential design planning processes that can be used in future safety evaluations that are conducted using health care database information.

VI. REFERENCES

1. Dal Pan GJ, Lindquist M, Gelperin K. Postmarketing spontaneous pharmacovigilance reporting systems. In: Strom BL, Kimmel SE, Hennessy S, eds. Pharmacoepidemiology. 5th ed: John Wiley & Sons; 2011:137-57.
2. Chen RT, Rastogi SC, Mullen JR, et al. The Vaccine Adverse Event Reporting System (VAERS). Vaccine 1994;12:542-50.
3. Rodriguez EM, Staffa JA, Graham DJ. The role of databases in drug postmarketing surveillance. Pharmacoepidemiol Drug Saf 2001;10:407-10.
4. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. N Engl J Med 2011;364:498-9.
5. Nguyen M, Ball R, Midthun K, Lieu TA. The Food and Drug Administration's Post-Licensure Rapid Immunization Safety Monitoring program: strengthening the federal vaccine safety enterprise. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:291-7.
6. Nelson JC, Cook AJ, Yu O, Zhao S, Jackson LA, Psaty BM. Methods for observational post-licensure medical product safety surveillance. Stat Methods Med Res 2015;24:177-93.
7. Robb MA, Racoosin JA, Sherman RE, et al. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:9-11.
8. Platt R, Archdeacon P, Bell C, et al. Mini-Sentinel Methods: Prospective Routine Observational Monitoring Program Tools (PROMPT) Users' Guide 2014 13 June 2014.
9. Brown JS, Kulldorff M, Petronis KR, et al. Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. Pharmacoepidemiol Drug Saf 2009;18:226-34.
10. Kulldorff M, Davis RL, Kolczak M, al. E. A maximized sequential probability ratio test for drug and vaccine safety surveillance. Seq Anal 2011;30:58-78.
11. Cook AJ, Wellman R, Marsh TL, et al. Statistical methods for estimating causal risk differences in the distributed data setting for postmarket safety outcomes: U.S. Fod and Drug Administration; 2012 March 1.
12. Cook AJ, Wellman RD, Nelson JC, Jackson LA, Tiwari RC. Group sequential method for observational data by using generalized estimating equations: application to Vaccine Safety Datalink. Journal of the Royal Statistical Society: Series C (Applied Statistics) 2014.
13. Taxonomy for monitoring methods within a medical product safety surveillance system: Year 1 report of the Mini-Sentinel Taxonomy Project Workgroup. 2010. (Accessed at http://www.mini-sentinel.org/methods/methods_development.)
14. Taxonomy for monitoring methods within a medical product safety surveillance system: Year 2 report of the Mini-Sentinel Taxonomy Workgroup. 2012. (Accessed at http://www.mini-sentinel.org/methods/methods_development)
15. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550-60.

16. Whitehead J. The design and analysis of sequential clinical trials. Revised, 2nd edition ed. Chichester, West Sussex: John Wiley & Sons Ltd; 2000.

17. Baggs J, Gee J, Lewis E, et al. The Vaccine Safety Datalink: a model for monitoring immunization safety. Pediatrics 2011;127 Suppl 1:S45-53.

18. Fireman B, Toh S, Butler MG, et al. A protocol for active surveillance of acute myocardial infarction in association with the use of a new antidiabetic pharmaceutical agent. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:282-90.

19. Page ES. Continuous inspection schemes. Biometrika 1954;41:100-14.

20. Wald A. Sequential tests of statistical hypotheses. Annals of Mathematical Statistics 1945;16:117-86.

21. Shih MC, Lai TL, Heyse JF, Chen J. Sequential generalized likelihood ratio tests for vaccine safety evaluation. Stat Med 2010;29:2698-708.

22. Nelson JC, Yu O, Dominguez-Islas CP, et al. Adapting group sequential methods to observational postlicensure vaccine safety surveillance: results of a pentavalent combination DTaP-IPV-Hib vaccine safety study. Am J Epidemiol 2013;177:131-41.

23. Li CI, Mathes RW, Bluhm EC, et al. Migraine history and breast cancer risk among postmenopausal women. J Clin Oncol 2010;28:1005-10.

24. Hocine MN, Musonda P, Andrews NJ, Farrington CP. Sequential case series for pharmacovigilance. Journal of the Royal Statistical Society series A 2009;172:213-36.

25. Cook AJ, Tiwari RC, Wellman RD, et al. Statistical approaches to group sequential monitoring of postmarket safety surveillance data: current state of the art for use in the Mini-Sentinel pilot. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:72-81.

26. Lan K, DeMets D. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-63.

27. Emerson S, Kittelson JM, Gillen D. Tutorial in biostatistics: Frequentist evaluation of group sequential clinical trial designs. Stat Med 2007;26:5047-80.

28. Guidance for industry: E9 statistical principles for clinical trials. 2015. (Accessed 20 July 2015, at http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf.)

29. Standards for the design, conduct, and evaluation of adaptive randomized clinical trials. 2015. (Accessed 20 July 2015, at http://www.pcori.org/assets/Standards-for-the-Design-Conduct-and-Evaluation-of-Adaptive-Randomized-Clinical-Trials.pdf.)

30. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. Epidemiology 2005;16:336-41.

31. Lieu TA, Kulldorff M, Davis RL, et al. Real-time vaccine safety surveillance for the early detection of adverse events. Med Care 2007;45:S89-95.

32. Klein NP, Fireman B, Yih WK, et al. Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. Pediatrics 2010;126:e1-8.

33. Yih WK, Nordin JD, Kulldorff M, et al. An assessment of the safety of adolescent and adult tetanus-diphtheria-acellular pertussis (Tdap) vaccine, using active surveillance for adverse events in the Vaccine Safety Datalink. Vaccine 2009;27:4257-62.

34. Belongia EA, Irving SA, Shui IM, et al. Real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. Pediatr Infect Dis J 2010;29:1-5.

35. Gee J, Naleway A, Shui I, et al. Monitoring the safety of quadrivalent human papillomavirus vaccine: findings from the Vaccine Safety Datalink. Vaccine 2011;29:8279-84.

36. Daley MF, Yih WK, Glanz JM, et al. Safety of diphtheria, tetanus, acellular pertussis and inactivated poliovirus (DTaP-IPV) vaccine. Vaccine 2014;32:3019-24.

37. Lee GM, Greene SK, Weintraub ES, et al. H1N1 and seasonal influenza vaccine safety in the vaccine safety datalink project. Am J Prev Med 2011;41:121-8.

38. Tse A, Tseng HF, Greene SK, Vellozzi C, Lee GM, Group VSDRCAIW. Signal identification and evaluation for risk of febrile seizures in children following trivalent inactivated influenza vaccine in the Vaccine Safety Datalink Project, 2010-2011. Vaccine 2012;30:2024-31.

39. Greene SK, Kulldorff M, Yin R, et al. Near real-time vaccine safety surveillance with partially accrued data. Pharmacoepidemiol Drug Saf 2011;20:583-90.

40. Silva IR, Kulldorff M. Continuous versus group sequential analysis for post-market drug and vaccine safety surveillance. Biometrics 2015.

41. Yih WK, Kulldorff M, Fireman BH, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. Pediatrics 2011;127 Suppl 1:S54-64.

42. Kulldorff M, Silva IR. Continuous post-market sequential safety surveillance with minimum events to signal. 2015.

43. Nelson JC, Cook AJ, Yu O, et al. Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using electronic observational health care data. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:62-71.

44. Zhao S, Cook AJ, Jackson LA, Nelson JC. Statistical performance of group sequential testing methods for post-licensure medical product safety surveillance: a simulation study. Statistics and its Inference 2012;5:381-90.

45. Tseng HF, Sy LS, Liu IL, et al. Postlicensure surveillance for pre-specified adverse events following the 13-valent pneumococcal conjugate vaccine in children. Vaccine 2013;31:2578-83.

46. Carnahan R, Gagne JJ, Nelson J, et al. MINI-SENTINEL SURVEILLANCE PLAN: MINI-SENTINEL PROSPECTIVE ROUTINE OBSERVATIONAL MONITORING PROGRAM TOOLS (PROMPT): RIVAROXABAN SURVEILLANCE. In: PROMPT Surveillance Plan; 2014.

47. Yih WK, Zichittella L, Sandhu SK, al. E. Accessing the freshest feasible data for conducting active Influenza vaccine safety surveillance; 2015.

48. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 1987;43:193-9.

49. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-56.

50. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191-9.

51. Kittelson JM, Emerson SS. A unifying family of group sequential test designs. Biometrics 1999;55:874-82.

52. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 1987;43:193-9.

53. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. Arch Intern Med 2012;172:1582-9.

54. Cook AJ, Wellman RD, Nelson JC, Jackson LA, Tiwari RC. Group sequential method for observational data incorporating confounder adjustment with application in postmarketing Vaccine Safety Datalink (VSD). Journal of the Royal Statistical Society 2015.

55. Cook AJ, Wellman RD, Marsh T, et al. Mini-Sentinel Methods Development: Statistical Methods for Estimating causal Risk Differences in the Distributed Data Setting for Postmarket Safety Outcomes: FDA's Sentinel Initiative: Project Report; 2012.

56. 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.

57. Maro JC, Brown JS, Kulldorff M. Medical product safety surveillance: how many databases to use? Epidemiology 2013;24:692-9.

58. Maro JC, Brown JS, Dal Pan GJ, Kulldorff M. Minimizing signal detection time in postmarket sequential analysis: balancing positive predictive value and sensitivity. Pharmacoepidemiol Drug Saf 2014;23:839-48.

VII. APPENDIX 1: CODE TO PERFORM SAMPLE SIZE ESTIMATION

The SAS code below (written in SAS software, version 9.3) includes two sample size calculation macros, one for the Regression method and one for the IPTW method, as well as examples of running these macros.

```
/*************************************************/
/* (1) Statistic = RR, Model = Logistic regression */
/*************************************************/
/* Sample size calculation for GS GEE method
/* Input parameters:
/*  OddsRatio = detectable odds ratio (e.g. oddsratio=2)
/*      Alpha = alpha level (e.g. alpha=0.05)
/*      Power = power (e.g. power=0.8)
/* Event_Rate = event rate among comparator group (e.g. event_rate=0.0004)
/* Pct_Expose = fraction of total sample in exposed group (e.g.
pct_expose=0.4) */
%macro seqRR(oddsratio,alpha,power,event_rate,pct_expose);
%let logOR=%sysfunc(log(&oddsratio));
%let beta=%sysevalf(1-&power);
%let weight=%sysevalf((1-&pct_expose)/&pct_expose);


proc seqdesign ALTREF=&logOR BOUNDARYSCALE=STDZ;
  /* Sample size calculations for different study designs */
  /* 1-time look */
  onetime: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=1;
  /* 4 looks, under 3 boundary shapes */
  Pocock_4x: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=4;
  InBetween_4x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.25) nstages=4;
  ObrienFleming_4x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.5) nstages=4;
  /* 8 looks, under 3 boundary shapes */
  Pocock_8x: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=8;
  InBetween_8x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.25) nstages=8;
  ObrienFleming_8x: design alpha=&alpha alt=UPPER beta=&beta method=POW
```

```
  /* 16 looks, under 3 boundary shapes */
  Pocock_16x: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=16;
  InBetween_16x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.25) nstages=16;
  ObrienFleming_16x: design alpha=&alpha alt=UPPER beta=&beta method=POW


  samplesize MODEL=TWOSAMPLEFREQ (TEST=LOGOR NULLPROP=&event_rate
REF=NULLPROP WEIGHT=&weight);
  /* Output boundaries and sample sizes for all designs */
  ods output samplesizesummary=ss method=boundary;
quit;
%mend seqRR;


/* Sample size calculations for OR=1.5, alpha=0.05, power=90%, event rate
among comparator group=0.000308, and 25% sample in exposed group */
%seqRR(oddsratio=1.5,alpha=0.05,power=0.9,event_rate=0.000308,pct_expose=0.25
);


/************************************************************/
/* (2) Statistic = Risk Difference, Model = Linear regression */
/************************************************************/
/* Sample size calculation for GS IPTW method
/* Input parameters:
/*   RiskDiff = detectable risk difference (e.g. riskdiff=0.0002)
/*      Alpha = alpha level (e.g. alpha=0.05)
/*      Power = power (e.g. power=0.8)
/* Event_Rate = event rate among comparator group (e.g. event_rate=0.0004)
/* Pct_Expose = fraction of total sample in exposed group (e.g.
pct_expose=0.4) */
%macro seqRD(riskdiff,alpha,power,event_rate,pct_expose);
%let beta=%sysevalf(1-&power);
%let weight=%sysevalf((1-&pct_expose)/&pct_expose);


proc seqdesign ALTREF=&riskdiff BOUNDARYSCALE=STDZ;
  /* Different study designs */
  /* 1-time look */
```

```
  onetime: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=1;

  /* 4 looks, under 3 boundary shapes */

  Pocock_4x: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=4;

  InBetween_4x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.25) nstages=4;

  ObrienFleming_4x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.5) nstages=4;

  /* 8 looks, under 3 boundary shapes */

  Pocock_8x: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=8;

  InBetween_8x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.25) nstages=8;

  ObrienFleming_8x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.5) nstages=8;

  /* 16 looks, under 3 boundary shapes */

  Pocock_16x: design alpha=&alpha alt=UPPER beta=&beta method=POW (RHO=0)
nstages=16;

  InBetween_16x: design alpha=&alpha alt=UPPER beta=&beta method=POW
(RHO=0.25) nstages=16;

  ObrienFleming_16x: design alpha=&alpha alt=UPPER beta=&beta method=POW



  samplesize MODEL=TWOSAMPLEFREQ (TEST=PROP NULLPROP=&event_rate REF=NULLPROP
WEIGHT=&weight);

  /* Output boundaries and sample sizes for all designs */

  ods output samplesizesummary=ss method=boundary;

quit;

%mend seqRD;


/* Sample size calculations for RD=0.00015, alpha=0.05, power=90%, event rate
among comparator group=0.000308, and 25% sample in exposed group */

%seqRD(riskdiff=0.00015,alpha=0.05,power=0.9,event_rate=0.000308,pct_expose=0
.25);
```

VIII. APPENDIX 2: DATA PULL USING CIDA TOOL

Individual-level data for the angioedema example were pulled at each Data Partner using the CIDA tool. All first episodes of medication use (defined in Section II above) regardless of episode length were pulled. Additional programs were written at Group Health and run at each Data Partner to create the 3 cohorts for the 4 analyses. Table A1 shows the input specifications in the CIDA tool query form for pulling the broader study cohort.

Table A1. CIDA tool input specifications for ACEIs inhibitors **and angioedema analyses**

CIDA TOOL INPUTS (numberings as in query form)	SPECIFICATION
1. Study Design	
Inclusion/exclusion criteria (this is just the summary, not input in query form)	Include: • Age 18+ years during Jan 2003-Dec 2012 • Members in Aetna, United, GH or KPNC • New users of ACE inhibitors, ARBs or BB defined as having 1 fill for the medication without any fills for any medications in the 3 classes of interest in the prior 183 days Exclude: • Concomitantly used medications in multiple therapeutic classes of interest on index date • History of angioedema (ICD-9 code 995.1) in any position during an outpatient, inpatient or ED visit in 183 days prior to index date
5. Query start date	1/1/2003
6. Query end date	12/31/2012
7. Age groups of interest	18-44, 45-54, 55-64 and 65+ years
8. Enrollment requirements	
Coverage type	Medical coverage: Yes (at index date and in the 183 days prior) Drug coverage: Yes (at index date and in the 183 days prior)
Maximum enrollment gap	45 days
Continuous enrollment before exposure	183 days
2. Exposures & Follow-up	
1. Exposures of interest	ACE inhibitors and ARBs
Comparator of interest	BB
2. Exposure incidence	
Wash-out period (number of days of continuous enrollment before the index date required to be free of the exposure)	183 days
Define exposure incidence	New users of ACE inhibitors, ARBs or BB defined as having 1 fill for the medication without any fills for any medications in the 3 classes of interest in the prior 183 days Exclude concomitantly used medications in

CIDA TOOL INPUTS (numberings as in query form)	SPECIFICATION
	multiple therapeutic classes of interest on index date (i.e., dispensed multiple drugs of interest on the index date)
3. Exposure time determined by	"Create treatment episodes" or use days supply associated with drug dispensings to create treatment episodes
4. When using days supply to determine exposure time	
Episode allowable gap (maximum number of days between 2 fills to create a single, continuous treatment episode)	14 days
Episode extension period (number of days to extend the length of a treatment episode)	0 day
Minimum episode duration (minimum treatment episode length to include)	0 day
Minimum days supply (to create treatment episodes)	0 day
5. Exposure time duration (number of days after treatment initiation to follow patients for the occurrence of the outcome)	Leave blank (N/A when using days supply to determine exposure time)
6. Number of exposure episodes allowed per person	"Retain the first episode only"
7. Exposure episode censoring rules (episodes are automatically censored at end of enrollment and the occurrence of outcome(s) of interest)	
Truncate exposure time when	Dispensation of a different therapeutic class of interest
8. Blackout period (exclude outcome if it occurred during this time following new medication use)	0 day
9. Additional inclusion criteria	None
10. Additional exclusion criteria	A history of angioedema (ICD-9 code 995.1) in any position during an outpatient, inpatient or ED visit in 183 days prior to index date
11. Sensitivity analyses	None
3. Outcomes	
1. Outcome of interest	Angioedema
2. Outcome definition	An ICD-9 code of 995.1 in any position during an outpatient, inpatient or ED encounter. Only the first occurrence is included
3. Outcome incidence	
Outcome washout	183 days
Additional criteria	None
4. Covariates	
1. Covariate evaluation window	183 days before index date
2. Covariate list (automatically included, n=12)	• Age • Sex • Time period • Year

CIDA TOOL INPUTS (numberings as in query form)	SPECIFICATION
	• Comorbidity score (not use in analysis) • # inpatient hospital stays (to be converted to a binary variable for 1+ stay in analytic dataset) • # non-acute institutional stay (not use) • # ED visits (not use) • # ambulatory visits (not use) • # other ambulatory visits (not use) • # drug dispensings (not use) • # unique generics dispensed (not use)
3. Covariate list (requester-defined)	Comorbid conditions: • Allergic reactions (ICD-9: 477.x, 518.6, 558.3, 691.x, 692.xx (except 692.75-692.77), 693.x, 708.x, 995.0, 995.27, 995.3, 995.6x, 995.7, V07.1, V13.81, V14.x, V15.0x, V72.7) • Diabetes (250.x) • Heart failure (402.x1, 404.x1, 404.x3, 428.xx) • Ischemic heart disease (410.x-414.x) Medication use (see drug list in Table 2): • NSAIDS • Aspirin • COX-2 inhibitors • Oral corticosteroids
Subgroups	None

IX. APPENDIX 3: ANALYTIC DATASET CREATION

Study Cohort for Each Analysis:

After the individual-level data were pulled at each Data Partner using the CIDA tool, additional programs were written at Group Health and run at each Data Partner to create the 3 cohorts for the 4 analyses. Table A2 shows the different specifications of the cohorts for the 4 analyses. Note that the cohorts for analyses #2 and #4 are identical.

Table A3.2. Study cohorts for the four ACE inhibitors and angioedema analyses

	Statistical Analyses			
	1 (GS GEE)	**2 (GS GEE)**	**3 (GS IPTW)**	**4 (GS IPTW)**
Index date	1st new use of ACE inhibitors, ARBs or BBs New use defined as no dispensings of any medications of interest in the 183 days before index)			
Range of follow-up[1] days from index	0-365	0-30	0-30	0-30
Min # days in treatment episode	0	0	30	0
Allow censoring events (disenrollment, end of treatment episode, switching to a different medication class) during follow-up	Yes	Yes	No	Yes
Cohort creation	Include all subjects Truncate follow-up time to 365 days	Include all subjects Truncate follow-up time to 30 days	Restrict to subjects with a minimum of 30 days treatment episode, and no censoring event during 30 days Truncate follow-up time to 30 days	Include all subjects Truncate follow-up time to 30 days

[1] Length of follow-up is defined as the number of days from index date until the earliest occurrence of the outcome, disenrollment from the health plan, cessation of the therapeutic class, and initiating another therapeutic class.

Cohorts for analyses 2 and 4 should be identical with the same number of subjects and outcomes.

Cohorts for analyses 1 and 2 (=4) should have the same number of subjects but analysis 1 cohort will have more outcomes than in the analysis 2 cohort because of its longer follow-up time.

Cohort for analysis 3 will have fewer subjects and outcomes than in the other 2 cohorts for analyses 1, 2 and 4.

Aggregated Datasets for GS GEE Method:

For Analyses 1 and 2 where the Poisson GS GEE method was applied, individual-level data were aggregated by categorical exposure status and confounder strata, whereby each row of the data set includes information about the frequency of select event(s) and the number of persons and person-time at risk in each exposure-confounder stratum.

Table A3.3 shows the structure of the aggregated dataset that were used in Analysis 2 when the maximum follow-up time is 30 days after the index date. A stratum is defined by all variables from DPID to S.

Table A3.3. Example aggregated dataset specifications for analysis 2 that will apply the Poisson GS GEE method on a cohort with a maximum of 30 days follow-up after the index date.

DPID	SITEID	AgeGroup	Sex	IndexYr	Confounders	X	S	N_Obs	Obs_t	Y_1	s_month	Looks$_i$	St$_i$	O$_{ij}$	C$_{ij}$	EO$_{ij}$	EC$_{ij}$	EA$_{ij}$
HM	GHC	18-44	F	2003	See below	0	1	100	32150	4								
HM	GHC	18-44	F	2003		1	1	35	9580	1								
HM	GHC	18-44	F	2003		2	1	20	7580	0								

Explanation of variable names in Table A3:

- **DPID:** Data Partner ID
- **SITEID:** Site ID
- **AgeGroup:** Age at index date (grouped: 18-44, 45-54, 55-64, 65+ years)
- **Sex:** F=Female, M=Male
- **IndexYr:** Year of index date (2003-2012)
- **Confounders**: Binary baseline confounders (in 183 days prior to index date)
  - **Allergic**: Allergic reactions (0=no, 1=yes)
  - **Diabetes**: Diabetes (0=no, 1=yes)
  - **Heart**: Heart failure (0=no, 1=yes)
  - **IHD**: Ischemic heart disease (0=no, 1=yes)
  - **AnyHosp**: Any inpatient hospitalization (0=no, 1=yes)
  - **NSAIDS**: Any traditional NSAIDS (0=no, 1=yes)
  - **Aspirin**: Any aspirin (0=no, 1=yes)
  - **COX2**: Any COX-2 inhibitors (0=no, 1=yes)
  - **OralCCS**: Any oral corticosteroids (0=no, 1=yes)
- **X:** Indicator variable of exposure of interest (0=BBs, 1=ACE inhibitors, 2=ARBs)
- **S**: Number of 84-day increments (integer) from study start date (1/1/2003) to index date. This variable will be used to adjust for calendar time in models
- **N_obs:** Total number of people in each stratum
- **Obs_t:** Total exposure time (contributed by N_obs people) in each stratum (in days)
- **Y_1**: Total number of outcome (occurred in N_obs people) in each stratum
- **S_month**: Month and year of study start
- **Looks, St, O, C, EO, EC** and **EA** represent arrays used to store summary information about follow up. These arrays and the method of aggregated and disaggregation are described below.

Summary of Data Aggregation for GSGEE (Macro: &sasmacr.data_ag_cohort.sas)

This section describes the method we developed and used to aggregate patient-level data for the angioedema data. It applies more broadly to any setting where individuals are followed for an outcome event, a censoring event (e.g. end of treatment episode) or the end of the study period (administrative censoring event). The method's purpose is to generate de-identified, aggregated data that can be used for analysis with the GS GEE method [54] in the case where follow up time varies in the study sample.

Tables A3.4 and A3.5 describe individual array elements and are followed by pseudo-code outlines of the algorithms.

Table A3.4. Details of SAS arrays used in data aggregation

NAME	TYPE/INDICES	DESCRIPTION
Looks(i)	1-D Array, *i = 1,..,L*, where *L* = the total number of analytic look periods.	Dates of the L analyses (looks) periods. Can also be thought of strictly as a set of periods that partition the study timeline.
St(i)	1-D Array, *i = 1,..,L*, where *L* = the total number of analytic look periods.	1/0 flag indicating whether the individual started the study during the i^{th} period. Value can be 1 for at most one value of *i*.
O(i,j)	2-D Array, *i,j = 1,..,L*, where *L* = the total number of analytic look periods.	1/0 flag indicating that the individual started during the period between look j and look *j-1*, and had an outcome in the period between look *i* and look *i-1*
C(i,j)	2-D Array, *i,j = 1,..,L*, where *L* = the total number of analytic look periods.	1/0 flag indicating that the individual started during the period between look j and look *j-1*, and was censored in the period between look *i* and look *i-1*, where censoring here is strictly defined as the end of a treatment episode.
EO(i,j)	2-D Array, *i,j = 1,..,L*, where *L* = the total number of analytic look periods.	The amount of time that the individual contributed to the i^{th} period given that they started in the j^{th} period and had an event in the i^{th} period. For each individual this will only be populated when j is equal to the period in which the individual started the study and i is the period in which the individual had an outcome event. Observed follow up time for the individual in periods prior to the one in which they had an outcome are collected in EA(*k,j*) where *k=j…i-1*.
EC(i,j)	2-D Array, *i,j = 1,..,L*, where *L* = the total number of analytic look periods.	The amount of time that the individual contributed to the i^{th} period given that they started in the j^{th} period and were censored in the i^{th} period. For each individual this will only be populated when j is equal to the period in which the individual started the study and *i* is the period in which the individual reached the end of a treatment episode. Observed follow up time for the individual in periods prior to the one in which they were censored are collected in EA(*k,j*) where *k=j…i-1*.
EA(i,j)	2-D Array, *i,j = 1,..,L*, where *L* = the total number of analytic look periods.	The amount of time that the individual contributed to the i^{th} period given that they started in the j^{th} period and were administratively censored at the end of the i^{th} period, i.e., their treatment episode continued into the next look period. For each individual this will only be populated when *j* is equal to the period in which the individual started the study and i>=j is not equal to a period in which the individual had an outcome event or reached the end of a treatment episode.

Table A3.5. Details of additional SAS arrays used in data expansion

NAME	TYPE/INDICES	DESCRIPTION
A(i,j)	2-D Array, i,j = 1,..,L, where L = the total number of analytic looks.	Aggregate count of administrative censorings at each analytic look period: A(i,j) = St(j) –(O(j,j)+…+O(i,j))- (C(j,j)+…+C(i,j)), for i>=j and where O(i,j), C(i,j) are the aggregate versions of the arrays defined in Table A4.
EOind(i,j)	2-D Array, i,j = 1,..,L, where L = the total number of	The average person time observed for individuals that entered the study in period j and had an outcome event in period i.

NAME	TYPE/INDICES	DESCRIPTION
	analytic looks.	EOind(i,j) = EO(i,j)/O(i,j) where EO(i,j) is the aggregate version of the array described in Table A4.
ECind(i,j)	2-D Array, i,j = 1,..,L, where L = the total number of analytic looks.	The average person time observed for individuals that entered the study in period j and had a censoring event in period i. ECind(i,j) = EC(i,j)/C(i,j) where EC(i,j) is the aggregate version of the array described in Table A4.
EAind(i,j)	2-D Array, i,j = 1,..,L, where L = the total number of analytic looks.	The average person time observed for individuals that entered the study in period j and had neither an outcome event nor a censoring event in period i. EAind(i,j) = EA(i,j)/A(i,j) where i>=j and A(i,j) = St(j) –(O(j,j)+…+O(i,j))- (C(j,j)+…+C(i,j)) where St(i), O(I,j), C(I,j) are the aggregate versions of the array described in Table A4.

DATA AGGREGATION ALGORITHM

1. Create global macro variables Increment, StudyStart, and StudyEnd; the time increment between looks, the study start date and the study end date, respectively.
2. Using patient-level data at DP, restrict to the years and variable categories specified in the study plan.
3. Create arrays as detailed in **Table A3.4**.
4. Compute variable Eventdt as the minimum of the outcome date and the episode end date.
5. Compute variable s_month, the year-month of each individual's indexdt, e.g., "200710".
6. If the individual's indexdt is within the study period, then compute the study period within which the indexdt falls.
7. Compute the final period to which the individual contributed person time.
8. Loop through each look period from 1 up to the current look period and do:
    a. Case 1: If the index value coincides with the individual's entry period:
        i. Check whether the index date and the event date are the same,
            1. If yes, then flag either O(i,Startlook) or C(i,Startlook) and allot 1 day to EO(i,Startlook) or EC(i,StartLook), respectively.
            2. If no, proceed to ii.
        ii. Take the minimum of Looks(i) and Eventdt and check whether an outcome or censoring occurred on or before the next scheduled analytic look:
            1. If yes, then flag either O(i,StartLook) or C(i,StartLook) and allot min(Looks(i),Eventdt) –indexdt to EO(i,StartLook) or EC(i,StartLook).
            2. If no, allot Looks(i)-indexdt to EA(i,StartLook).
    b. Case 2: If the index value is in between the entry period and the last period for the individual:
        i. Allot Looks(i)-Looks(i-1) to EA(i,StartLook)
    c. Case 3: If the index value coincides with the individual's final study period:
        i. Check whether an outcome or censoring occurs in the indvidual's final study period

1. If yes, flag O(i,StartLook) or C(i,StartLook) and, if the index is not equal to StartLook (handled in case a.) Allot Eventdt-Looks(i-1) to either EO(i,Startlook) or EC(I,StartLook).
2. If no, allot Looks(i) – Looks(i-1) to EA(i,StartLook).

9. Aggregate data by covariate levels and s_month, summing over all values of the following arrays: St(i),O(i,j), C(i,j),EO(i,j),EC(i,j),EA(i,j).

DATA EXPANSION ALGORITHM (Performed once data is returned to MSOC)

1. Initialize and assign global macro variables Increment, StudyStart, StudyEnd, Num_increments, n_var; the study period increment, study start date, study end date, total number of increments during the study period, number of variables needed in arrays.
2. Beginning with the aggregate data, designate arrays for the aggregate counts and exposure times per **Table A3.4**.
3. Initialize arrays for new computed variables as given in **Table A3.5**.
4. Begin a nested loop on two dimensions: 1) over looks from 1 to the current look by 1, and 2) over start looks from 1 to CL by 1.
    a. For each (i,j), i>=j, sum outcomes and censorings from (j,j) to (i,j), i.e., for a given look and start look, count the number of individuals that had an outcome or a censoring (episode end) from the time they started through the current look. The difference between the number that started in a particular look period and the count above is the number of administrative censorings that would occur at the current look, A(i,j). A(i,j) is the number of administrative censorings at look i that started at look j.
    b. For each (i,j) compute the following
        i. EOind(i,j) = EO(i,j)/O(i,j)
        ii. ECind(i,j) = EC(i,j)/C(i,j)
        iii. EAind(i,j) = EA(i,j)/A(i,j)
5. Begin nested loop on two dimensions: 1) for *l* from the current look to 1 by -1, and 2) *m* from *l* to 1 by -1.
    a. Check whether the two indices are equal, i.e., *m* = *l*.
        i. If yes,
            1. Check O(*l,m*) > 0.
                a. If yes, loop from 1 to O(*l,m*)
                    i. Y = 1
                    ii. Obs_t = EOind(*l,m*)
                    iii. Output
            2. Check C(*l,m*) > 0.
                a. If yes, loop from 1 to C(*l,m*)
                    i. Y = 0
                    ii. Obs_t = ECind(*l,m*)
                    iii. Output
            3. Check A(*l,m*) > 0 & l = CL.

a. If yes, loop from 1 to A(*l,m*)
    i. Y = 0
    ii. Obs_t = EAind(*l,m*)
    iii. Output

b. Check whether m < l.
    i. If yes,
        1. Check O(m,l) > 0.
            a. If yes, loop from 1 to O(l,m)
                i. Y = 1
                ii. Obs_t = EAind(m,m)+…+EAind(l-1,m) + EOind(l,m)
                iii. Output
        2. Check C(m,l) > 0.
            a. If yes, loop from 1 to C(l,m)
                i. Y = 0
                ii. Obs_t = EAind(m,m)+…+EAind(l-1,m) + ECind(l,m)
                iii. Output
        3. Check A(m,l) > 0 & l=CL.
            a. If yes, loop from 1 to A(l,m)
                i. Y = 0
                ii. Obs_t = EAind(m,m)+…+EAind(l,m)
                iii. Output