MINI-SENTINEL METHODS

COLLECTING SUPPLEMENTAL INFORMATION VIA TWO-PHASE STUDY DESIGNS TO INVESTIGATE SIGNALS ARISING FROM MEDICATION SAFETY SURVEILLANCE ACTIVITIES

Prepared by: Sascha Dublin, MD, PhD,[1,4] Rod Walker, MS,[1] Carolyn Rutter, PhD,[1,4] Jennifer Nelson, PhD,[1,4] Bruce Fireman, MA,[2] David Graham, MD, MPH,[3] Bruce Psaty, MD, PhD,[1,4] Soko Setoguchi, MD, DrPH,[5] and Azadeh Shaoibi, MS, MHS[3]

Author Affiliations: 1. Group Health Research Institute, Seattle, WA; 2. Kaiser Permanente Northern California, Oakland, CA; 3. Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD; 4. University of Washington, Seattle, WA; 5. Duke Clinical Research Institute, Durham, NC.

December 18, 2013

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.

Acknowledgments

The authors would like to thank the other members of this Mini-Sentinel workgroup, listed below, for their participation. Their contributions have been invaluable.

The authors would also like to thank Katherine Yih, PhD, MPH, of Harvard Pilgrim Health Care Institute; Elizabeth Chrischilles, MS, PhD, of The University of Iowa College of Public Health; and Michael Nguyen, MD, and Scott K. Winiecki, MD, of FDA's Center for Biologics Evaluation and Research, for providing additional input.

Mini-Sentinel Operations Center staff based at Harvard Pilgrim Health Care Institute, including Kara Coughlin, Susan Forrow, Jillian Lauer, and Nicholas Lehman-White, provided project support. We thank Candace Fuller of the Operations Center for her thoughtful review of a draft.

Table 1. Additional Members of the Mini-Sentinel Supplemental Information for Improved Confounder Adjustment Workgroup

Affiliation*	Name
HealthCore, Inc., Wilmington, DE	Dan Mines, MD, MSCE
Kaiser Permanente Georgia, Atlanta, Georgia	Robert Davis, MD, MPH
University of Pennsylvania Perelman School of Medicine, Philadelphia, PA	Kevin Haynes, PharmD, MSCE
Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, MD	Erika Avila-Tang, PhD, MHS Manuel Bayona, MD, MS, PhD Yelizaveta Torosyan, MD, PhD
Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD	Aloka Chakravarty, PhD Eric Frimpong, PhD, MA Brad McEvoy, DrPH Yu-te Wu, PhD
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA	Darren Toh, ScD

*Affiliation at the time of workgroup participation.

Mini-Sentinel

Mini-Sentinel Methods

Collecting Supplemental Information via Two-Phase Study Designs To Investigate Signals Arising From Medication Safety Surveillance Activities

Table of Contents

I. EXECUTIVE SUMMARY

The Food and Drug Administration's Mini-Sentinel program conducts active safety surveillance for a wide range of medical products and outcomes. When a signal emerges, there may be concern that the administrative data used for routine surveillance are not sufficient to account for confounding, either because data elements of interest are not included or because available measures lack accuracy. In either case, supplemental data collection may be warranted, e.g., through medical record review. Two-phase study designs can improve the efficiency of supplemental data collection and can also provide additional information about outcomes or exposures. In many settings, two-phase studies can substantially reduce bias while maintaining acceptable precision, thus helping to resolve some of the uncertainty that will be present when signals arise from routine surveillance activities. This report provides guidance about the use of two-phase study designs within Mini-Sentinel. The primary focus is on collecting supplemental confounder data, because that was the workgroup's charge, but we also discuss scenarios where there is a need to collect supplemental data about exposures or outcomes.

Two-phase study designs use the information available on the whole population (the "phase 1 data") to identify the most informative people to target for supplemental data collection. The supplemental information, called the "phase 2 data", is then used to estimate the magnitude of the signal while more fully adjusting for confounders. This report includes general background about two-phase studies as well as guidance about study design, analytic approaches, and practical considerations. In most contexts it will be desirable to stratify the population on phase 1 measures of exposure and outcome, then sample an equal number of people from each stratum for supplemental data collection ("a balanced design"; see page 7). Decisions must also be made about the phase 2 sample size, and for this a simulation-based approach can be helpful, as outlined in the final section of this report.

Two-phase analytic techniques address two issues: 1) the need to account for differential selection of patients from the phase 1 sample into the phase 2 sample, to avoid bias; and 2) the use of information from the phase 1 sample to improve precision. Several methods are available, and more research is needed to clarify how well they perform, especially with small phase 2 sample sizes. Conducting a two-phase study raises many practical considerations. We describe questions that need to be answered to design a two-phase study (p. 52, Table 8) and explore them by working through an example. It is important to consider whether supplemental data are needed for the outcome or exposure, as well as confounders. This choice will affect the resources required and by implication, the phase 2 sample size. Also one must decide at which Data Partners (DPs) to collect supplemental data. We recommend targeting DPs in whose data the signal was observed as well as those contributing a large proportion of exposed person-time and outcome events. We considered a prospective approach – that is, collecting supplemental information while initial surveillance activities are underway – but until a signal has arisen, it is difficult to ensure that data collection will target only the most informative subjects.

The final section of this report presents a simulation-based approach to answering study design questions. Simulations can help determine how beneficial a two-phase study might be in a given context and the phase 2 sample size that would likely be needed. We have developed tools for conducting such simulations which will be available to all Mini-Sentinel teams. In conclusion, the use of two-phase study designs is a promising approach to improve the efficiency of supplemental data collection that may be needed when a signal emerges from Mini-Sentinel surveillance activities.

II. INTRODUCTION AND CONTEXT

In 2008, the FDA launched the Sentinel Initiative aiming to create a new national system for monitoring medical product safety. The ultimate goal is to utilize real-world health care data to allow early detection of any safety problems associated with newly introduced medical products including drugs and devices. When a safety signal emerges, the FDA and its partners will need to investigate further to determine whether the signal reflects a true safety problem or is a spurious result caused by an alternative process, such as bias from confounding.

A spurious signal could arise due to confounding when one or more characteristics are associated both with the medical product under study and the outcome of interest but are not adequately controlled for in the analysis of the exposure/outcome relationship. One reason for inadequate control of confounding could be that the electronic data sources include only poorly measured proxies for some important confounders, or in some cases, no information at all.

This report focuses on one strategy for addressing confounding in such scenarios: the two-phase study design. Such a design entails using the information available on everyone (referred to as the "phase 1 data") to identify a targeted subgroup for collection of supplemental confounder information from alternative sources such as medical records. The supplemental information, called the "phase 2 data", is then incorporated into analyses using two-phase analytic methods. In this introduction, we will describe the context in which such supplemental data collection would be carried out, including the signal evaluation activities that would precede it. We will briefly discuss how the two-phase design relates to alternative approaches that are not examined in this report. We will also describe the contents and structure of the report which follows.

When a safety signal arises from prospective surveillance, the FDA and its partners will embark upon further investigation. Many steps will be carried out before the team considers collecting supplemental information, such as from medical charts, because such new data collection is costly and time consuming. First, the team will conduct initial data checks and signal exploration, as outlined in a previous report.[1,2] Examples of actions at this stage include looking to see if the signal is present only at one or a few DPs or sub-sites; checking for programming errors; and examining the temporal pattern of outcomes. If concern remains, additional steps will be taken using existing data within the Mini-Sentinel Common Data Model (MSCDM). Such steps may include additional analyses which adjust for a richer and more customized set of confounders. They may also include quantitative bias analyses which explore the potential impact of different biases upon the results and consider the magnitude and direction of bias that would need to be present to explain the observed results. If concern persists about the signal after these initial steps have been completed, the next step (which will be the focus of this report) is to consider whether sufficient clarification of the signal could be obtained by collecting more detailed data via a two-phase study.

This report focuses on the use of supplemental data collection to improve control of confounding. There may be concern about other potential biases as well, such as outcome misclassification and protopathic bias (that is, unclear temporal relationship between the exposure and the outcome.) There may be a desire to collect supplemental data about the outcome and/or exposure. While those activities are not the focus of this report, two-phase study designs can be useful for targeting supplemental data collection in those contexts as well, and much of this report is applicable to those situations.

In this report, we focus primarily on medical record review as a means of collecting supplemental data about confounders, but other approaches could be used, such as mailing questionnaires, conducting telephone interviews, carrying out physical measurements or obtaining genomic data. These approaches may not be well suited to many active surveillance activities as they are more expensive, time consuming, and burdensome to subjects than medical record review. The methods we discuss in this report related to two-phase study design and analysis are just as applicable in those scenarios, since all of these modes of data collection will benefit from an approach that targets only the most informative people for data collection. In addition to study design and analysis, this report also addresses practical considerations, including how to target chart review or other data collection when there are multiple DPs representing a large number of health care systems and providers.

A key aspect of the two-phase design strategy described here is that supplemental data collection targets a subgroup of subjects chosen because of characteristics that can be measured in the existing electronic data. There are several motivations for such targeting, including 1) increased efficiency and 2) avoiding selection bias. We recognize that in some cases, there may be additional electronic data available from a subset of Data Partners beyond what is in the Common Data Model – for instance, laboratory values or vital signs. It may be useful to take advantage of these existing data instead of collecting new data. Such a scenario – which we will refer to as "opportunistic" supplemental data collection – has some overlap with the scenario that is our main focus, but there are important differences. This report will not address all of the challenges related to using such opportunistic data. Instead, the report which follows will focus on the deliberate sampling scenario, and only this approach will be referred to using the terminology of the statistical literature as a "two-phase study." In a later section of this report, we will highlight the similarities and differences between two-phase study designs and opportunistic supplemental data collection, and we will point out what additional methodological work is needed to make the best use of opportunistic data.

Three sections follow this Introduction. In Section III, we focus on methodologic considerations related to two-phase study designs. We discuss the background, context and uses for such studies, then discuss design considerations. Next we describe two-phase analytic techniques that can be used to incorporate such data into analyses. Section IV focuses on the practical and logistical aspects of designing a two-phase study within Mini-Sentinel. In this section, we work through a detailed example to illustrate questions that must be addressed and lay out challenges and suggested solutions. Finally, Section V focuses on how simulation studies can help answer particular design questions, including whether a two-phase study is likely to be helpful in a specific scenario. We present results of a simulation study based on the example from Section IV and focus in particular on the effect of different phase 2 sample sizes on power and bias reduction.

III. DESIGN AND ANALYSIS OF TWO-PHASE STUDIES

This section of the report reviews the goals, design and analysis of two-phase studies from a methodological perspective. Practical and logistical considerations are presented in Section IV, which follows this section.

We begin this section by defining a two-phase study, noting scenarios that lend themselves to use of a two-phase design. Next, we review issues related to the two-phase study design, such as stratification of the phase 1 data and sample selection. Finally, we review analytic approaches for two-phase studies.

A. TWO-PHASE STUDIES

Two-phase studies are used to estimate the association between an exposure and outcome when:

1) An initial (phase 1) sample has been collected and outcome and exposure information are known for the entire sample; and
2) Additional important information can be collected for a selected subsample (phase 2). This additional information could include new or more accurate information about potential confounders or more accurate information about the outcome and exposure.

Two-phase studies are used to obtain ***unbiased estimates*** of an exposure-outcome relationship using phase 2 data, which contain detailed and accurate outcome, exposure, and confounder measures, in conjunction with less detailed phase 1 data which contain information about how the phase 2 sample was selected (i.e., the probability of selection into the phase 2 sample). The goal in designing a two-phase study is to develop a phase 2 sampling plan – a way of choosing patients – that provides the most information for a given sample size. That is, ***the phase 2 sample is selected so that it results in the greatest precision for a given sample size***.

1. Study settings

Several study settings may give rise to two -phase data, and we describe four such settings below. Setting 1 will be the primary focus of this report; however, aspects of the other settings will be discussed.

For simplicity, in this report we focus on a dichotomous exposure and a dichotomous outcome, with their association measured by the log odds ratio, which may be estimated using logistic regression models. We use the following notation:

X	True exposure (dichotomous, X=1 exposed, X=0 unexposed)
Y	True outcome (dichotomous, Y=1 present, Y=0 absent)
β	Log odds ratio measuring the X-Y relationship: $\ln\left(\frac{\Pr(X=1\mid Y=1)}{\Pr(X=0\mid Y=1)}\right)-\ln\left(\frac{\Pr(X=1\mid Y=0)}{\Pr(X=0\mid Y=0)}\right)$
Z_1	Confounders available from the phase 1 data
Z_2	Confounders available from the phase 2 data
X'	Error prone exposure measure, available from phase 1 data
Y'	Error prone outcome measure, available from phase 1 data

a. Setting 1

The simplest setting for a two-phase study occurs when exposure (X) and outcome (Y) information are available *without error* for the entire study sample (i.e., the phase 1 data), but important confounder information is not available except through sampling a subset of the population for more intensive data collection (i.e., the phase 2 data).[3] **In this setting, the primary reason for conducting a two-phase study is to collect richer and more accurate confounder information.** This scenario was the requested focus of the current task order and is relevant to the Mini-Sentinel setting because many confounders are difficult to measure using administrative data alone.

For example, smoking status and obesity may be important confounders in some Mini-Sentinel surveillance activities. While International Classification of Diseases, version 9 (ICD-9) codes for tobacco use and obesity may be available within routine Mini-Sentinel data (imprecise 'phase 1' confounders: Z_1), the information that could be obtained through medical record review is expected to be far more complete and accurate (more accurate 'phase 2' confounders: Z_2). There may also be additional confounders of interest (such as disease severity) not routinely available in administrative (phase 1) data which are only available through additional (phase 2) manual medical record review.

b. Setting 2

Another setting for two-phase studies occurs when an error prone exposure measure is available from the phase 1 data (X'), but the true exposure (X) is only available from phase 2 data. This setting could arise within the Mini-Sentinel project in the case of exposures such as biologic agents or devices, for which procedure codes in administrative data may not contain adequate detail. More information may be desired, e.g., manufacturer name, lot number, characteristics of the product such as concentration, etc. **In this setting, the goal of conducting a two-phase study is to gather additional information about confounders and to get more accurate exposure information.**

This setting has been commonly described in the literature about two-phase studies, mostly focusing on exposures other than medications or devices. Often, the phase 1 data arise from a case-control study (when outcome status is the basis for selection into the phase 1 sample), and the second phase sample is used to obtain both confounder information and gold standard exposure information.[4-7] This setting, with a phase 1 case-control study, is the context used for development of many of the design and analysis methods used for two-phase studies. In general, this situation is less likely to occur within Mini-Sentinel because, in practice, computerized pharmacy data are considered the "gold standard" for measuring exposure to prescription medications, and it is unlikely that supplemental data collection of any kind would be able to improve on the phase 1 measures of exposure in most cases. However, there are scenarios relevant to Mini-Sentinel in which exposure measurement could be a concern. For example, when examining the safety of devices or biological agents, administrative data may provide inadequate information about the product, and it may be desirable to gather more detailed data from medical charts, such as the specific manufacturer of a device.

c. Setting 3

A third setting for two-phase studies occurs when an error prone *outcome* measure is available from the phase 1 data (Y'), but the true outcome (Y) is only available from phase 2 data. **In this situation, the goal of new data collection is to gather additional information about confounders and to validate the outcome.** This setting is also relevant to Mini-Sentinel activities because many outcomes of interest are poorly measured in administrative data. Although this setting has not been as commonly addressed in

the existing statistical literature, it can be addressed in the same manner as Setting 2, assuming that only the outcome is measured with error. In drug safety surveillance, there is great variability across outcomes in terms of how accurately they can be measured from administrative data. Some outcomes (such as myocardial infarction) are believed to be measured with high accuracy from administrative data. While error in measuring outcomes at phase 1 may be a reality for some Mini-Sentinel activities, or even a major concern, this scenario will not be a focus of the current report. As discussed above, the primary charge of this workgroup was to explore methods to improve confounder measurement. Thus we focus primarily on the scenarios in which improved confounding control is of highest priority in understanding a potential signal that has arisen during routine safety surveillance.

d. Setting 4

The setting least frequently considered in the literature occurs when both the exposure and outcome are potentially misclassified so that error prone exposure and outcome data are available from the phase 1 data (X', Y'), but the true exposure (X) and outcome (Y) data are only available from the phase 2 sample. **In this scenario, the motivation for conducting a two-phase study is to collect confounder information and to collect better information about both the exposure and the outcome of interest.** As in setting 2 above, this scenario would be uncommon within Mini-Sentinel because the phase 1 exposure measure based on pharmacy data is likely considered the "gold standard". Little research has been carried out to address design and analysis issues when both the outcome and exposure are measured with error, and discussion of this context falls outside the scope of the current report.

In all four of these settings, the general approach of the two-phase study is the same:

1) First stratify the phase 1 sample into groups on the basis of the information known for everyone; and

2) Then sample individuals from each stratum to form the phase 2 dataset.

The choice of the phase 2 sampling plan affects the precision of the estimated exposure-outcome relationship. Therefore, the goal is to strategically select the phase 2 subjects to maximize precision (minimize variability) of the estimated association of interest.

2. Stratification and sampling

To help orient the reader to the stratification and sampling steps of a two-phase study, we begin by considering the simplest setting when both exposure and outcome information are assumed to be available without error for the phase 1 data. Figure 1 on the next page (adapted from Collet, Schaubel, Hanley et al., 1998[3]) demonstrates a simple two-phase design, assuming that the dichotomous outcome and exposure are measured without error and that no additional confounder information is available in the phase 1 data. In this scenario, confounder data will be obtained for a phase 2 sample that is drawn from the stratified phase 1 sample, with patients stratified by exposure and disease outcome.

There are four basic ways to select the phase 2 sample: 1) patients can be randomly sampled without regard to exposure or outcome status; 2) patients can be sampled based on outcome only (i.e., a case-control sample can be selected); 3) patients can be selected based on exposure only; or 4) patients can be selected based on *both* outcome and exposure. Random sampling of patients is useful when both the exposure and outcome are common. This scenario is unlikely to arise in Mini-Sentinel surveillance activities. Outcome-based sampling is useful when the outcome is rare, but the exposure is common.

Figure 1: Simple Two-phase Design

Phase 1

		Disease: Yes	Disease: No
Drug Exposure	Yes	N_1	N_2
	No	N_3	N_4

Phase 2

		Disease: Yes	Disease: No
Drug Exposure	Yes	n_1	n_2
	No	n_3	n_4

Exposure-based sampling is useful when the exposure is rare, but the outcome is common. It is necessary to sample based on both the exposure and outcome when both the exposure and outcome are rare. This is the scenario we expect to encounter most commonly within Mini-Sentinel activities.

After deciding to select the phase 2 sample from groups defined by exposure and outcome, one must then decide how many to sample from each group. A balanced design selects an equal number of patients within each of the phase 1 strata. Several papers[5,7,8] have shown that a *balanced design* is an efficient sampling approach for a two-phase study. In other words, a balanced design is expected to result in an estimated log odds ratio (measuring the exposure-outcome relationship) that has the best precision (smallest variability) for a given phase 2 sample size. Under a balanced design, the selection probabilities vary across strata. For example, the probability of selecting a patient with both drug exposure and the outcome of interest is equal to n_1/N_1. Patients in small phase 1 strata have a higher probability of selection into the phase 2 sample. The process of stratification and oversampling of patients from small strata is used to improve the efficiency of estimated effects of exposure on the outcome. Patients are not selected with equal probability, and failing to adjust for this would result in selection bias. However, because the selection mechanism is known (i.e., is by design), analyses used to estimate the association between exposure and outcome can adjust for and remove the induced selection bias.

The above illustration assumed only exposure and outcome information were known at phase 1 (i.e., it did not consider additional confounder information available at phase 1). There may be scenarios when additional confounder information is available from the phase 1 data. For example, in Mini-Sentinel, ICD-9 codes available at phase 1 may provide important confounder information. In such settings, these confounders could be used to further stratify the phase 1 data (i.e., stratification by outcome, exposure, and confounder) in an effort to further improve the precision of two-phase estimates.

3. Two published examples that use simulation to demonstrate the utility of two-phase sampling

Hanley and Dendurki (subsequently referred to as H&D) illustrate the use of two-phase sampling using a *simulation study* based on a study designed to estimate the risk of upper gastrointestinal (GI) hemorrhage associated with use of selective serotonin reuptake inhibitor antidepressants (SSRIs).[9] In their example, exposure (SSRI use) and outcome (upper GI hemorrhage) information are available for 44,199 patients in the phase 1 sample, shown in Table 2. At phase 2, a subset of 1,000 patients is selected. The phase 2 data includes key confounder data such as body mass index (BMI), smoking history, and heavy alcohol use (all combined, for simplicity, into a 0/1 dichotomized confounder score, C, to represent background risk).

H&D demonstrate the impact of different methods of selecting the phase 2 sample on the standard error of the estimated association between SSRI use and upper GI hemorrhage (based on the log odds ratio) and show the efficiency gains possible with a two-phase study that selects patients based on

Table 2: Phase 1 data for the Hanley & Dendurki (2009) example

		Upper GI hemorrhage		
		Yes	No	Total
SSRI use	Yes	335	1,780	2,115
	No	3,693	38,391	42,084
	Total	4,028	40,171	44,199

both exposure and outcome relative to other sampling approaches. (We discuss the specific analytic approaches used to remove the selection bias induced by the two-phase methods in Section III.C.1. below.) Each of the phase 2 sampling schemes can be analyzed to yield (asymptotically) unbiased estimates of the true adjusted association of interest; therefore, comparisons of sampling schemes focus on comparisons of standard errors (i.e., comparisons of *efficiency* of estimation).

Because this is a simulation study, H&D actually knew the confounder information, C, for all 44,199 patients, allowing them to compare results of their two-phase sampling designs to the 'truth' based on an analysis that adjusted for C in the entire (phase 1) population. H&D used the entire phase 1 dataset to estimate the association between SSRI use and upper GI hemorrhage first without any confounder information, resulting in an estimated odds ratio, exp($\hat{\beta}$), equal to 1.96 with a standard error for $\hat{\beta}$ equal to 0.062. Then they showed the estimated association between SSRI use and upper GI hemorrhage adjusting for C in the entire phase 1 dataset, resulting in an estimated odds ratio equal to 1.72, with standard error equal to 0.063. The unadjusted estimate is what one would get without undertaking a two-phase study (and thus would be biased for the true adjusted association of interest), and the adjusted estimate is the best one could do (best possible precision, i.e., smallest standard error) if C were available for the entire phase 1 data. In an actual data analysis, you would not know the smallest possible standard deviation, but this provides a reference for their simulated two-phase analyses.

Table 3 summarizes the range of standard error estimates for second-phase sampling designs examined by H&D in their simulations. This simulation study shows that a balanced two-phase sampling design in which only 1,000 patients are selected for key confounder ascertainment yields standard error estimates that are only slightly larger than those that would have been obtained had confounders been ascertained for all 44,199 patients in the phase 1 sample and had analyses been based on this much larger dataset.

Table 3: Summary of simulation results comparing different two-phase sampling designs, from Hanley & Dendurki (2009)

Design	Range of standard error estimates
Entire Sample (44,199 patients)	0.063
1: Random sample of 1000	0.39 to 0.72
2: Case-Control 500 cases, 500 controls	0.26 to 0.33
3: Exposed-Unexposed: 500 exposed, 500 unexposed	0.20 to 0.22
4: Balanced: 250 from each exposure/outcome strata	0.079 to 0.092

As H&D note, the key to gaining efficiency via two-phase designs is selection of patients to provide the most information possible. This can be seen in their example by looking again at the distribution of the phase 1 data (Table 2) and comparing this to the sampling probabilities (Table 4, next page) and the

average number of patients selected from each strata under the four competing two-phase designs (shown in Table 5).

Table 4: Second phase sampling probabilities of each exposure-outcome strata for each two-phase sampling design, from Hanley & Dendurki (2009)

	User Case	User Control	Non-user Case	Non-user Control
1: Random Sample	0.02	0.02	0.02	0.02
2: Case-Control	0.12	0.01	0.12	0.01
3: Exposed-Unexposed	0.24	0.24	0.01	0.01
4: Balanced	0.75	0.14	0.07	0.01

Table 5: Average number of patients selected from each exposure-outcome strata for each two-phase sampling design, from Hanley & Dendurki (2009)

	User Case	User Control	Non-user Case	Non-user Control	Total
Population	335	1,780	3,693	38,391	44,199
1: Random Sample	8	40	83	869	1,000
2: Case-Control	42	22	458	478	1,000
3: Exposed-Unexposed	79	421	44	456	1,000
4: Balanced	250	250	250	250	1,000

Less than one percent of the entire population is an SSRI user with upper GI hemorrhage, yet this is the most informative patient group, in terms of precise estimation of the association between this exposure and outcome. Because of this, only the design scheme that samples on the basis of both exposure and outcome succeeds in obtaining good capture of this group.

Collet and colleagues (1998)[3] provide another illustrative example of two-phase sampling in which they examine the relationship between low birth weight and preschool asthma using a cohort of 16,207 children. As in the previous example, confounder data are available for the full cohort (5 binary confounder variables). They use these data to compare the results of a complete cohort analysis (with confounder data measured on all study subjects with the results of a two-phase analysis where confounder data were only available for a selected phase 2 sample of 400 children. (Once again we discuss the specific analytic approaches used to remove the selection bias induced by the two-phase methods in Section III. C.1.) Of the competing phase 2 sampling approaches considered (random, case-control, balanced), their results show the balanced approach does best in regard to efficiency of the estimated association between low birth weight and preschool asthma, as well as efficiency of the estimated associations between the confounders and risk of asthma. As in the prior example, of the designs considered, the balanced design results in the most efficient estimates because of the small number of phase 1 patients in the exposed case strata (only 58 children have both low birth weight and preschool asthma). This is reflected by the average number of children selected from each exposure-outcome strata, shown in Table 6 (next page).

Table 6: Average number of children selected from each exposure-outcome strata for each two-phase sampling design of 400 selected children (from Collet et al.[3]), and the corresponding standard error estimate of the log odds of the low birth weight – preschool asthma association

	Exposed Case	Exposed Control	Unexposed Case	Unexposed Control	Standard error
Population	58	170	1,438	14,541	0.20
1: Random Sample	1	4	36	359	*
2: Case-Control	8	2	192	198	0.67
3: Balanced	58	114	114	114	0.45

*Results not presented by Collet et al.[3]

The prior two illustrative examples reveal that even in the simplest two-phase setting (known dichotomous outcome and exposure, and no additional confounder information available in the phase 1 data), efficiency gains are highly dependent on the design of the phase 2 sample.

Efficiency gains in more complex settings where additional strata could be formed on the basis of available phase 1 confounder data will also rely heavily on phase 2 design choices. While the above examples point to a balanced design, more efficient designs may exist depending on the true relationships between exposures, outcomes, confounders, and surrogate measures of confounders (about which we may have little information). The existence of such optimal designs could be explored via a simulation study, as we discuss later on in Section III.B.2.c.

This report will focus on two-phase sampling when patients are sampled based on both outcome and exposure status and potentially additional available confounder data available at phase 1. It will address the *design* of the phase 2 sample (e.g., what proportion of patients should be sampled from each strata to maximize efficiency) and statistical methods for *analysis* of these data to obtain unbiased estimates.

B. TWO-PHASE STUDY DESIGN

The two-phase study design has three basic components: ***sample size*** at both phase 1 and phase 2; ***stratification*** of phase 1 data; and ***phase 2 selection probabilities*** that vary across the phase 1 strata. This combination of phase 1 stratification and phase 2 selection probabilities that vary across phase 1 strata is a mechanism for targeting patients for more detailed data collection with the goal of maximizing precision (minimizing variability) of the estimated association between exposure (X) and outcome (Y). This area of research has focused on one goal: minimizing the variance of the estimated log odds ratio, β, describing the association between X and Y for a fixed phase 2 sample size.

The choice of phase 1 strata and phase 2 selection probabilities directly affects the precision of the estimated exposure-outcome relationship. Therefore, the key questions to be addressed when developing a two-phase design are:

1. How should patients be stratified using phase 1 data?
2. How should patients be selected from phase 1 strata to optimize efficiency? That is, what selection probabilities should be used to sample patients from the phase 1 strata to minimize the variance of the estimated exposure-outcome relationship?
3. How many total patients should be sampled at the second phase?

4. What additional data should be collected at the second phase? What principles should guide this decision?

We address each of these questions in turn below, focusing on contexts likely to be relevant to the Mini-Sentinel project.

1. How should patients be stratified using phase 1 data?

The choice of phase 1 strata depends on the study settings (described in Section III.A.1.), that is, it depends on what is known for the entire (phase 1) sample. Our review focuses on Setting 1, when both exposure (X) and outcome (Y) information are assumed to be available without error in the phase 1 data, but some important confounder information is not available except through additional data collection. We anticipate that this will be the most likely setting for the use of two-phase designs within Mini-Sentinel.

When both X and Y are known without error, phase 1 data should be stratified, at a minimum, on the basis of both X and Y. For a given sample size, 'more efficient' study designs are designs that result in more precise estimation of the X-Y association. As the examples in Section III.A.3. demonstrate, stratifying on both X and Y can be a more efficient sampling design than stratifying on the basis of only one of these variables. Furthermore, stratifying on both X and Y will not be less efficient than stratifying on just one of these variables. Efficiency gains are greatest when both the exposure (X) and outcome (Y) are rare. In Mini-Sentinel we often will be conducting surveillance for relatively rare outcomes, and rarity of the exposure will depend on the drugs and populations chosen for surveillance activities.

In addition to exposure and outcome information, covariate data are often available in phase 1 data. This additional data, Z_1, may be well measured confounders of interest (e.g., demographics, other medication usage, etc.), or they may be error prone or surrogate measures of confounders of interest (e.g., administrative codes for comorbid conditions) that could be ascertained without error in a phase 2 sample (Z_2). If covariate data, Z_1, are available at phase 1, then the efficiency of the estimated X-Y association of interest may be improved by stratifying the phase 1 data on the basis of exposure (X), outcome (Y), and the additional phase 1 data (Z_1). Efficiency gains from additional stratification on Z_1, relative to stratifying only on X and Y, are dependent on the associations between X, Y, Z_1, and Z_2. Therefore, when deciding how to stratify the phase 1 data, it is important to understand the expected relationships among these variables.

For example, suppose that study investigators believe that the presence of comorbid illnesses (Z_2) may confound the hypothesized exposure-outcome (X-Y) association, but that accurate measurement of Z_2 is only possible through phase 2 data collection. If age (Z_1) is available at phase 1, and older age is associated with presence of comorbid illness, then one approach to improving efficiency is to stratify the phase 1 data on the basis of exposure (X), outcome (Y), and age groups (Z_1). This stratification will allow targeted sampling of older age groups that are more likely to have the comorbid conditions that may confound the X-Y association. Efficiency gains will depend on various factors such as: how strong a prognostic factor comorbid illness Z_2 is for the outcome Y; how strongly Z_2 is associated with exposure X; and, the prevalence of Z_2 and magnitude of its association with Z_1. For example, if comorbid illness is actually fairly common across the population or not strongly associated with age (especially after stratifying on exposure), then targeting phase 2 data collection to certain age groups will not have as much potential for efficiency gains.

Key point: Stratifying on a phase 1 covariate (in addition to stratifying on the primary exposure and outcome) may result in efficiency gains when:

1) A somewhat rare but important confounder of interest is to be collected at phase 2, AND
2) The phase 1 covariate is strongly associated with the phase 2 confounder of interest.

There are no hard-and-fast answers, though, in terms of how rare a confounder would need to be or how strong the association would need to be for researchers to see efficiency gains with such a two-phase sampling design. Answers will depend on the particular study setting.

In the Mini-Sentinel study examining the association between saxagliptin (X) and acute myocardial infarction (MI, Y), which is discussed in detail in Section IV of this report, both exposure and outcome are available from phase 1 administrative data. These phase 1 data also provide error-prone measures of confounders, including obesity and smoking. For the saxagliptin study, phase 1 strata could be formed on the basis of saxagliptin exposure status, MI outcomes in the 12 months following initiation of saxagliptin or a 'control' medication, and administrative data measures of obesity and/or smoking during the baseline period (e.g., 12 months prior to the initiation of a study medication). Selecting subjects for phase 2 data collection based on phase 1 strata formed using these administrative data measures could have multiple efficiency benefits. In addition to allowing a more targeted sampling of likely smokers or obese subjects for gold standard data collection about smoking and BMI, these strata may also be associated with other important confounding conditions such as diabetes severity. Thus, sampling from strata based on the imperfectly measured phase 1 confounder data could provide better opportunity to select the informative people with severe diabetes complications at phase 2. It is important to note, however, that there is a limit to the number of phase 1 strata that can be formed. At some point the strata could become so small that two-phase analytic methods could yield unstable estimates of the adjusted X-Y association of interest.

Key conclusion: Phase 1 strata should always be formed on the basis of both exposure and outcome. The feasibility and benefit of more complex designs will depend on factors such as sample size and the relationships present in a particular setting. Simulation studies can provide insight into these issues, as we discuss in Section III.B.2.c. below and explore in detail in Section V of this report.

a. Note on other settings

We recognize that phase 1 Mini-Sentinel data may be prone to errors in observed exposure or outcome variables, and we address stratification for these settings below.

Studies focusing on medication exposure (X) often rely on records of prescriptions filled, but this information does not necessarily translate to actual usage, and administrative data codes for some biologic agents or medical devices may not be adequate to provide all of the information desired about an agent (e.g., manufacturer name, lot number, characteristics of the product such as concentration, administered dose, etc.). Several papers provide guidance for phase 1 stratification when an error-prone exposure measure *X'* is available at phase 1 and true exposure can only be obtained at phase 2 data collection (our "Setting 2," described in Section III.A.1.b.) In this setting, it is "never a disadvantage" to stratify the phase 1 data using both the known outcome (Y) and the error-prone exposure measure (*X'*).[4] That is, study designs based on stratification on Y and *X'* are at least as efficient as designs that stratify on Y alone. The efficiency gains from additional stratification on *X'*, relative to stratifying on Y alone, depend on the associations between X and both *X'* and Y. Efficiency gains increase when *X'* is more

accurate (less error-prone) and when there are stronger associations between outcome status and exposure. This discussion assumes that a gold standard exposure measure is available from more intensive phase 2 data collection. As noted in the Mini-Sentinel context, however, in many cases the primary medication exposure available from administrative data is regarded as the "gold standard," which is why this setting is not discussed in detail in this report.

In regard to outcomes (Y) in Mini-Sentinel, it is important to understand that some outcomes are much better measured through administrative data codes than others. For example, myocardial infarction is coded with relatively high sensitivity and positive predictive value (e.g., PPV 86% in a recent study carried out within Mini-Sentinel[10]),[11] and it is highly likely that the date of the administrative code corresponds closely to the date of the actual event. In contrast, outcomes such as acute liver or renal failure are captured less accurately, and in some cases, the date of onset may not correspond well to the onset date determined using administrative data. Thus a potential study scenario is one in which exposure X and an error-prone outcome measure *Y'* are known at phase 1, but the true outcome can only be captured at phase 2 data collection (our "Setting 3," described in Section III.A.1.c.) In this case, the same general principal as above applies. The phase 1 data should be stratified, at a minimum, on the basis of the available exposure and outcome information (i.e. both X and *Y'*). We note that there is ongoing work being done by the Protocol Core as part of the Prospective Routine Observation Monitoring Program Tools (PROMPT) activity which is geared toward identifying and prioritizing outcomes with the most accurate administrative algorithms for routine surveillance. This is designed to minimize (to the extent possible) outcome misclassification; as such, this setting is also not discussed in detail in this report.

2. How should patients be selected from phase 1 strata to optimize efficiency?

Once the phase 1 sample has been stratified, the next step is to choose sampling fractions that will be used to select patients from these strata. For each stratum, the sampling fraction is equal to the number selected for phase 2 data collection divided by the total number of phase 1 patients in the stratum. The choice of sampling fractions (i.e., how many patients to select from each stratum) will affect efficiency in estimating β, the log odds ratio describing the association between X and Y. Sampling fractions that result in smaller expected standard errors of the estimate of β are considered better, i.e., more efficient designs. We have already discussed that random sampling or case-control sampling is not as efficient as a balanced design, i.e., one in which sampling fractions are chosen to yield an equal number of patients from each of the phase 1 strata. Thus, for this section we will discuss balanced and optimal designs, and the factors that govern efficiency.

a. Balanced and optimal designs

A **balanced design** selects an equal number of patients from each stratum. If the number of patients available in the phase 2 data in any stratum is smaller than the planned phase 2 sample size, then all patients in the stratum are selected with the remaining phase 2 sample size equally distributed across the remaining strata. This approach effectively oversamples rare events, increasing the expected number of patients in uncommon strata. The balanced design is extremely useful because it is generally an efficient design, though it may not be the optimal design.

The **optimal design** is one that results in the smallest possible expected standard error of the estimate of β (i.e., it is the most efficient design). Optimal designs may have different sampling fractions from a

balanced design and these sampling fractions are dependent on many factors, including the prevalence of and the relationships between the exposure, outcome, confounders, and stratification variables. As a hypothetical example, for one setting the optimal design might be to select every case regardless of exposure status, along with an unbalanced number of controls across strata defined by exposure and possibly other phase 1 covariates. Determining the optimal design for a particular setting, however, requires additional information, which may not be available at the time of design. Further, depending on the setting, the optimal design may only be marginally better (in terms of efficiency) than a balanced design.

Key point: Balanced designs perform well in most contexts and do not require assumptions about parameter values. Results for optimal designs are based on asymptotic efficiency, and estimation error may reduce the potential gains in efficiency for optimal designs. Optimal designs require, at a minimum, very good prior information or pilot data from the target population. Thus, the utility of optimal designs is highly dependent on the study setting. In cases where optimization is needed, simulation likely provides the best approach for determining the value of optimization. Such simulations are discussed in Section III.B.2.c. below.

b. Factors that affect efficiency

When both the exposure (X) and outcome (Y) are available without error in phase 1, but important confounder information is not available except through additional phase 2 data collection (Z_2), the efficiency of the balanced design and the phase 2 sampling fractions that correspond to the optimal design both depend on exposure, outcome, and confounder prevalence, and the odds ratios (ORs) measuring associations between X and Y (OR_{XY}), Z_2 and Y (OR_{ZY}), and X and Z_2 (OR_{XZ}). To choose effectively between study designs, and to decide when an optimal design might be necessary, it is important to understand scenarios that can lead to loss of efficiency.

Collet et al (1998)[3] used a simulation study to investigate factors affecting the efficiency of a balanced two-phase study, where phase 1 strata are defined by outcome (Y) and exposure only (X), and phase 2 data are collected on a confounder (Z_2). They assumed phase 1 data were from a case-control study with 1000 cases and 2000 controls. Their 'base case' setting assumed 20% exposure prevalence, 15% confounder prevalence, OR_{XY}=1.5, OR_{ZY}=3.0, and OR_{XZ}=3.0. They simulated selection of a phase 2 sample of size 500 using a balanced design and computed estimates of $\beta(=\ln(OR_{XY}))$ and its standard error. They repeated this simulation, each time varying one of the factors (e.g., decreasing exposure prevalence, increasing confounder prevalence, strengthening OR_{XZ}, etc.).

These simulations found that efficiency of the balanced design was poorest when confounder prevalence was set very high (>90%) or very low (<10%). Efficiency of the balanced design also decreased as the magnitude of the exposure-confounder association (OR_{XZ}) or the confounder-outcome association (OR_{ZY}) were increased—that is, with stronger confounding. Varying the exposure prevalence in these simulations had less influence on efficiency than varying confounder prevalence because the phase 1 data were stratified by exposure. Also, perhaps not surprisingly, efficiency increased when Collet et al simulated data with more cases available at phase 1 (e.g., higher outcome prevalence) or simulated a larger phase 2 sample.

One limitation of these simulations is that they assume very strong associations between confounders and both the outcome (OR_{ZY}=3.0) and the exposure (OR_{XZ}=3.0), at least relative to the exposure-

outcome association (OR_{XY}=1.5). Specifically they assume confounder associations that are double the magnitude of the assumed exposure-outcome association. Whether such relatively strong confounder relationships would be anticipated within Mini-Sentinel likely depends on the exposure-outcome association and population under study.

Breslow and Cain (1988)[5] examined efficiencies of phase 2 sampling designs in a rare disease setting. Like the Collet et al example above, they did this in a context where they assumed that exposure X and outcome Y data were available on all subjects at phase 1, with phase 2 data collection being performed to collect an additional confounder Z_2. In their example, X, Y and Z_2 are all dichotomous, and they estimate the X-Y association using a logistic regression model:

$$\log(pr(Y=1)/pr(Y=0)) = \beta_0 + \beta_1*X + \beta_2*Z_2$$

To relate this to the Mini-Sentinel saxagliptin example, one can think of saxagliptin use as X, MI occurrence as Y, and Z_2 as an important confounder such as smoking status measurable only at phase 2. Breslow and Cain then compared the efficiency of various phase 2 sampling plans for estimating β_1 (e.g., the association between saxagliptin use and MI occurrence adjusted for smoking status). They compared a case-control design that selected an equal number of cases and controls (stratify by Y only), a balanced design that selected an equal number of patients in each X-Y strata, and an optimal design that used sampling fractions specifically chosen to provide the most efficient estimate of β_1 for each simulated scenario (e.g., an assumed strength of the associations, β_1 and β_2, and the prevalence of Y). These optimal sampling designs were determined using a grid search. That is, rather than using a formula to calculate optimal sampling fractions, these were identified empirically, by systematically evaluating the efficiency of different sampling fractions based on the simulated scenario, and selecting those found to give the minimum variance for estimating β_1. While these simulations provide a useful reference, they do not provide a practical design approach, because they use information that could only be known without error in the context of a simulation study (e.g., the true strength of the confounder effect and the true exposure-confounder association among controls).

Breslow and Cain (1988)[5] showed that the balanced design was considerably more efficient than the case-control design when the association between Z_2 and Y was strong (i.e., large β_2), and was approximately as efficient as the case-control design when there was no association between Z and Y (i.e., β_2=0). While it may seem unlikely that investigators would carry out a two-phase study when β_2=0 (i.e., when Z_2 actually is not a confounder), it is important to consider analytic approaches that will work well under all plausible circumstances. When comparing the balanced design to the optimal design (for a given simulated scenario), Breslow and Cain found that the balanced design *was optimal* when there was no association between X and Y (i.e., β_1=0) and *was near optimal* when the number of cases was small relative to the number of controls (i.e., when the outcome of interest was rare). The reason behind this is that in such a scenario, both the balanced and optimal designs would essentially select all of the cases. Schaubel, Hanley, Collet et al (1997)[8] point to Woolf's formula for the variance of the log of the OR[12] as the rationale for why the balanced approach is the most efficient. However, Breslow and Cain also found that the balanced design could be considerably less efficient than the optimal design for estimating β_1 when there was a strong association between X and Y (i.e., large β_1) along with large numbers of both cases and controls available in the phase 1 sample.

c. Simulations to assess utility of a two-phase study

In light of the relationship between often unknown factors and optimal designs, Haneuse, Schildcrout, and Gillen (2012)[13] advocate a simulation-based approach to two-phase study design that considers the potential impact of confounding on efficiency. They focused on power, that is, the ability to detect an association between X and Y (OR_{XY}). Power increases with efficiency, that is, as the standard error of log OR_{XY} decreases. Haneuse, Schildcrout, and Gillen recommend computing bounds for power (assuming a fixed sample size) across a range of plausible (assumed) values for the exposure-outcome relationship (OR_{XY}) and confounding associations (OR_{ZY}, and OR_{XZ}). These can then be used to determine whether the desired power is likely to be achieved with a particular study design. While the framework and goals of their paper are not exactly the same as the context of Mini-Sentinel, the principles that they set forth are applicable. Once a signal is generated from phase 1 Mini-Sentinel data, a small simulation study could be performed prior to phase 2 data collection to establish power bounds that determine the utility of going forward with a phase 2 study.

Suppose that for the Mini-Sentinel saxagliptin study, phase 1 strata are formed on the basis of saxagliptin use (X) and MI occurrence (Y), with the goal of measuring the presence of BMI and smoking (Z_2) at phase 2 and then adjusting for these confounders when estimating the exposure-outcome association (OR_{XY}). Because it is unrealistic that study investigators will know the associations between all of these factors prior to study completion or the bias that exists in an unadjusted vs. adjusted analysis, simulation studies can be used to estimate power of a two-phase study design under various assumptions, and the results of these can be used to determine the potential utility of performing a two-phase study. Such a simulation becomes more complex, with more assumptions necessary, if the phase 1 sample is further stratified by error-prone administrative data measures of BMI and smoking categories available in phase 1 data (Z_1) in addition to stratification on saxagliptin use (X) and MI occurrence (Y). This increased complexity occurs because now, the efficiency of phase 2 sampling fractions depends on assumed relationships with Z_1 in addition to the other odds ratios and exposure, outcome, and confounder prevalence. Thus, a simulation has many more combinations of assumed inputs to consider.

To carry out a simulation that assesses utility of a two-phase study design in the context of a Mini-Sentinel study that has generated a signal from phase 1 data, an investigator must make reasonably accurate educated guesses about the possible relationships between the phase 1 variables and the phase 2 variables to be collected. When parameters defining these relationships are uncertain, simulation studies may result in a broad range of possible results. One approach is to first carry out a small pilot study to assist in study design. However, such repeated chart review may not be possible within the context of Mini-Sentinel.

Key conclusions: Determining the utility of a two-phase study and the relative benefits of competing designs is likely best accomplished via a simulation study. This is due to the fact that the performance of a two-phase study design is influenced by the prevalence of exposure, outcome, and confounders; the strength of associations between the exposure, outcome, and confounders; and phase 1 and 2 sample sizes. Simulation studies can provide insight into the relative influence of these factors, both in general and tailored to specific examples. In regard to choosing a balanced vs. a potentially optimal design, the balanced design has the benefit of being simple and easy to implement and describe, and it avoids major inefficiencies that occur with random or case- or exposure-only stratification. In contrast, optimal designs can only be developed and implemented if the investigator assumes knowledge about factors

such as the true strength of the confounder effect and the true exposure-confounder association among controls. Simulation studies offer a way to compare possible efficiency gains from optimal designs (relative to a balanced design) by hypothesizing a plausible range of settings (i.e., varying the assumed truth of the outcome, exposure, and confounder prevalence and the odds ratios OR_{XY}, OR_{ZY}, OR_{XZ}). That is, simulation can be used to determine how much an optimal design might improve the precision of estimates compared to a balanced design when the assumptions made in the simulation are actually observed in practice. Simulation can also be used to help determine how much an optimal design can worsen precision if the assumptions made in the simulation turn out to be incorrect.

In Section V of this report we provide results from a limited simulation study, based on a Mini-Sentinel surveillance activity, that investigates the influence of some of the factors and design choices discussed above on the performance of a hypothetical two-phase study.

3. How many patients, total, should be sampled at the second phase?

The number of patients selected at the phase 2 sample will depend on multiple factors, including the phase 1 sample design, the strength of the association between outcome and exposure (OR_{XY}), the desired precision (or power), and the resources available.

When both the exposure (X) and outcome (Y) are available without error in phase 1 data and a balanced design is used for the phase 2 sample, sample size (or power for a given sample size) depends on the exposure and confounder prevalence and their association, OR_{XY}.[8] When a case-control study is used at phase 1, exposure attributes (OR_{XY} and prevalence) have the greatest impact on power for a given sample size (assuming a balanced design). As noted earlier, determining the sample size needed to achieve a certain power using a balanced design that has phase 1 strata constructed on the basis of exposure, outcome, and additional known phase 1 covariates (confounders, surrogates, etc.) is challenging because sample size estimation requires assumptions about relationships between the additional covariates and other variables (e.g., exposure and outcome).

Based on the work of Haneuse, Schildcrout, and Gillen (2012),[13] simulation-based power calculations that examine a range of assumptions are likely to give the most realistic information about power that can be achieved within a two-phase study. This would be an especially important approach to conduct if there is a known upper bound on the phase 2 sample size, as simulation based power calculations using that maximum sample size could help determine whether there is any chance of achieving the level of precision desired. This would help determine whether phase 2 data collection will be worth the resources required (including time and money).

4. What additional data should be collected at the second phase?

Decisions about phase 2 data collection are largely driven by scientific issues, that is, an understanding of what is needed to provide valid answers to the question of whether a signal is true. For example, it is essential to understand what confounders really matter and therefore need to be measured at phase 2. Investigators must think about both the strength and prevalence of potential confounders. For a rare confounder to have a meaningful impact on the estimated association between the exposure and outcome, it needs to also be a strong confounder. This balance between the prevalence and strength of a confounder is an important consideration as investigators decide which confounders to target in two-

phase data collection. Furthermore, it is important to avoid introducing ascertainment bias during phase 2 data collection; this can be avoided by seeking the same information for all sampled patients.

Several logistical aspects also must be considered when planning phase 2 data collection. A basic consideration is whether the required data are even available from more intensive sampling. Monetary costs are another important consideration. The greater the per patient cost of phase 2 data collection, the fewer the total patients that can be sampled. The time period required for data collection also must be considered. Some of these logistical aspects within the context of Mini-Sentinel may lead investigators planning a two-phase study to focus on only a few sites where the important confounders can actually be obtained or where costs might be less prohibitive. If such an approach were taken, however, questions regarding generalizability must be addressed.

Ultimately, there are no simple answers to what data should be collected, as this will depend on the specific confounders, the ease of data collection, the expected size of the signal, the degree of confounding, and the overall budget. A more detailed discussion of these issues will be provided in Section IV of this report.

C. ANALYSIS OF TWO-PHASE STUDIES

When analyzing two-phase data, there are two essential issues that must be dealt with. First, the analyst must avoid bias when estimating the association between exposure and outcome by accounting for differential selection of patients from the phase 1 sample into the phase 2 sample. Second, the analyst should use information available from the phase 1 sample to improve the precision of estimates.

The specific model used to address these two issues while estimating the association between the exposure and outcome in a two-phase study depends on the outcome of interest. For example, a logistic regression model would be used for dichotomous outcomes and a survival model would be used for time to event outcomes. Throughout this document, we have assumed a dichotomous outcome with estimation of the exposure effect based on logistic regression. Our discussion of efficient study design focused on minimizing the variance of the estimated log odds ratio, $\hat{\beta}$, describing the association between X and Y for a fixed phase 2 sample size. In this section, we provide further details about estimation of logistic regression models when data are collected using a two-phase design.

1. Estimation of logistic regression models using two-phase samples

The data collected in a two-phase design consist of the phase 1 data and the phase 2 data, both of which are assumed to be random samples of patient-level variables that depend on the true associations between variables in the underlying population, as well as the sampling design used to collect the phase 1 data (e.g., case-control design, prospective cohort, etc.). Fitting a logistic regression model to the data requires specifying a *likelihood,* which is a statistical term that refers to the probability of observed data conditional on unknown model parameters (e.g., coefficients in the logistic regression model). The overall data likelihood associated with a two-phase design is the product of two components: the likelihood associated with the phase 1 sample and the likelihood associated with the phase 2 sample, conditional on phase 1 data. The likelihood is a function of both observed data and unknown model parameters. Estimating the values of the parameters involves a maximization of this likelihood function given the observed data. When data are collected using a two-phase design this process is complicated because both phase 1 and phase 2 data likelihoods need to account for whether data are collected

prospectively (i.e., drawn from a cohort) or were collected retrospectively (i.e., based on observed data, as in a case-control study). Retrospective data collection induces *constraints* on the likelihood.[6,14] That is, sample sizes within each phase 2 strata are fixed by design rather than varying freely. Similarly, if the phase 1 sample is from a case control study, then the number of cases and controls are fixed.

Various approaches have been proposed for estimating coefficients in a logistic regression model in the context of a two-phase design. There are three basic estimation approaches based on different formulations of the likelihood: weighted likelihood, pseudo- or profile likelihood, and maximum likelihood. Essentially, these approaches differ with respect to how the likelihood is defined and thus how the method combines information from the phase 1 and phase 2 samples, and the degree to which the likelihood incorporates constraints on the parameter space.

While all the methods yield consistent (i.e., asymptotically unbiased) estimates of the odds ratio, the efficiency of estimation (i.e., size of the standard errors) can vary across the different likelihood based estimation methods. Further, the performance of these methods in the presence of small samples, assumption violations, or model misspecification can differ. Overviews of these three estimation methods and comparisons of efficiency in the context of logistic regression analysis are provided by Breslow and Holubkov (1997),[6] Breslow and Chatterjee (1999),[7] and Haneuse, Saegusa, and Lumley (2012)[15]. We summarize some of this information below.

The ***weighted likelihood*** approach is conceptually simple; it modifies the likelihood of the phase 2 data by weighting observations based on the phase 1 sampling fractions,[16] also known as a Horwitz-Thompson estimator (1952).[17] This approach is also related to the method described by White (1982)[18] who recommended weighting observed stratum frequencies by sampling fractions, and using these weighted frequencies to estimate odds ratios. The weighted likelihood approach extends this idea to the logistic regression context. The idea behind the weighted likelihood is simple. One keeps track of the sampling fractions used to select subjects from the phase 1 data for phase 2 data collection. Then, observations are reweighted (using these known sampling fractions) in the calculation of the likelihood so that the sample is representative of the target population. Thus, weighting accounts for differential selection of patients from the phase 1 sample into the phase 2 sample so that the analyses avoid bias when estimating the association between exposure and outcome, and information about the phase 1 sample is incorporated through these weights.

However, the weighted likelihood approach ignores the constraints induced by two-phase sample design. These constraints induce correlation in the phase 2 sample, because observations are not independently selected. Standard errors are estimated using a robust covariance estimator that accounts for the correlation induced by the two-phase sampling approach (also known as the 'sandwich estimator').[19] This robust covariance estimator modifies the parametric covariance estimator, based on the information matrix, with logistic regression scores that incorporate empirical variability. The use of the sandwich estimator is known to have little effect on the variance estimates unless the dataset is very small or the logistic regression model is 'grossly misspecified'.[15]

The ***pseudo-likelihood*** and ***profile likelihood*** approaches account for constraints induced by the phase 1 sample design, but ignore the phase 2 constraints. The pseudo-likelihood estimation approach of Breslow and Cain (1988),[5] originally referred to as a conditional maximum likelihood, is similar to weighting but differs with respect to how weights are incorporated. This estimation approach conditions

on second-phase selection probabilities and treats these as known quantities that are incorporated into the binomial likelihood underlying the logistic regression model via *offset* terms. An offset is a special covariate in a regression model. The coefficient parameter for an offset is not estimated, rather it is fixed at one. In Breslow and Cain's approach, a logistic regression model is fit to the phase 2 data with each observation assigned a stratum-specific offset equal to the log of the odds of selection into the second phase sample (the odds being the probability of selection divided by the probability of not being selected). Valid variance and covariance estimates for model parameters relating to the strata can be obtained by applying simple corrections (based on the strata sizes) to the covariance matrix. Schill et al. (1993)[20] propose a slightly modified pseudo-likelihood approach that also utilizes stratum-specific offsets (different ones than the Breslow and Cain approach) but in a logistic regression model fit jointly to the first and second phase data. As such, their method uses a different correction to the variance and covariance estimates. Scott and Wild (1997)[21] developed a related ***profile likelihood*** approach that treats the second phase selection probabilities as unknown parameters. The profile likelihood approach iteratively estimates unknown parameters by first conditioning on probabilities (treating them as known) while estimating logistic regression parameters, then conditioning on logistic regression parameters (treating them as known) to re-estimate probabilities (and so on). Lee, Scott & Wild (2010)[22] claim the profile likelihood estimation method is more efficient than the pseudo-likelihood approach of Breslow and Cain and that in some cases there can be big differences in efficiency.

Breslow and Holubkov (1997)[14] and Scott and Wild (1991,1997)[23,24] also developed estimators for logistic regression models of two-phase data using a full ***maximum likelihood*** approach, which accounts for constraints induced by both phase 1 and phase 2 sampling. Breslow and Holubkov provide an algorithm for the implementation of this maximum likelihood estimation approach which entails an iterative fitting of a series of logistic regression models. Their simulation results also demonstrate improved efficiency relative to the other two-phase estimation approaches under certain data settings. The improved performance results from fully incorporating sample size constraints imposed by the two-phase design.

2. Comparison of estimation methods

Overall, the different approaches provide similar estimates. Simulation studies by Breslow and Holubkov (1997A & B)[6,14] find that in most cases results from maximum likelihood and profile likelihood methods are very similar. In an extreme case, when the stratification and explanatory variables were strongly associated, maximum likelihood was more efficient than profile likelihood. Both maximum likelihood and profile likelihood tend to be more efficient than weighted likelihood. However, weighted likelihood tends to be more robust to misspecification of the logistic regression model than either maximum likelihood or profile likelihood.

Given the likely relatively small sizes of Mini-Sentinel phase-two samples, it would be useful to have a better understanding of differences in small-sample properties of the three estimation methods across various design choices, as well as the potential impact that assumption violations or model misspecification might have on estimation. Toward this end, we point to the usefulness of a simulation based approach to investigating these issues for a particular study context. The manuscript and corresponding R package developed by Haneuse, Saegusa, and Lumley (2012)[15] are designed to support this type of investigation.

3. Available software

Two-phase methods are part of a larger body of survey sampling methods, and methods for correctly analyzing these data, including correct variance estimation, are widely available in software packages designed for complex survey sampling. We outline a few of these below; however, when utilizing any of these software packages (or any statistical software), it is imperative to consult the current software documentation and understand the underlying estimation procedures and associated assumptions.

SAS provides procedures, including SURVEYLOGISTIC, that can be used to analyze two-phase data. This procedure estimates logistic regression models using a pseudo-likelihood approach with standard errors estimated, by default, using Taylor series expansion; however, other estimation options may be specified. SAS also has procedures that enable estimation of proportional hazards models using survey samples.

The 'survey' package in R provides a 'svyglm' function that can be used for estimation. This function fits a generalized linear model to data from complex survey designs, with inverse probability weighting and design-based standard errors. This package also has a 'twophase' function geared specifically for the two-phase design. The 'osDesign' package in R is useful for planning a two-phase study as it allows for the evaluation of performance of various logistic regression estimation methods within the context of different two-phase design choices. R packages can be found at http://cran.r-project.org.

Like R, Splus has complex survey sampling routines. Stata also provides a suite of survey sampling routines that can be used to analyze two-phase sample data.

4. Estimation of survival models using two-phase samples

The literature on two-phase designs focuses on retrospective studies: that is, study designs that stratify a larger sample based on observed information after outcomes have already occurred. Reflecting this retrospective design, the two-phase literature focuses almost exclusively on analysis of a dichotomous outcome and methods for using logistic regression models to estimate the association between an exposure and the outcome. A few articles have considered the use of survival models when data are collected using a two-phase design.[25-28] Each of these articles assumes that data collection has been carried out using a usual retrospective two-phase design, based on a dichotomous outcome, and discusses survival analysis in the context of analytic methods to incorporate data from the phase 2 sample into analyses. This is a relatively recent area of statistical research, and research published to date focuses on the statistical properties of estimators and test statistics, not design issues associated with selection of the two-phase sample.

When data are obtained using a two-phase sampling design, weighted partial likelihood methods can be used to obtain unbiased estimates of regression parameters associated with a Cox proportional hazards models, and these methods are widely available in software used to analyze complex survey data. Breslow and Wellner (2007)[28] showed that Cox regression parameters have an asymptotic Normal distribution enabling the use of Wald tests for statistical significance. This means that as the sample size becomes very large ("asymptotically"), a test of H_0: $\beta = 0$ can be based on comparison of $\hat{\beta} / \text{se}(\hat{\beta})$ to a standard normal distribution, where $\text{se}(\hat{\beta})$ is the standard error of $\hat{\beta}$. The Wald test is based on a Taylor series expansion of the likelihood ratio test and while these two tests are asymptotically

equivalent (i.e., equivalent as sample sizes become very large), likelihood ratio tests are generally preferred to Wald tests because they are more reliable in small samples.[29] However, the chi-square comparator distribution (the distribution of the test statistic under the null hypothesis) used for likelihood ratio tests assumes simple random sampling and therefore requires modification when it is used in the context of a two-phase design. Formulae for the modified likelihood ratio statistics are available for computation of the appropriate comparator distribution, resulting in a modified log-likelihood statistic, but this approach does not appear to be implemented in survey sampling analysis software available from either SAS or Stata.

To our knowledge, no published papers have considered design issues related to stratifying and sampling patients for two-phase studies based on time-to-event outcomes. As noted previously, the two-phase design is used to direct data collection efforts to the most informative patients. Two-phase studies are especially useful when both the outcome and exposure are rare and typically oversample patients with both the exposure and outcome. Extending the idea of targeted data collection to time-to-event outcomes requires additional knowledge or assumptions about which patients are most informative, e.g., whether outcomes that occur early on are equally or more informative as those occurring later. These decisions depend on both the underlying hazard function (i.e., how risk changes with time) and clinical and/or biological knowledge about the exposure-outcome relationship. An example of such knowledge might be information about whether events occurring later on are as likely to be causally related to exposure (based on biologic or physiologic processes) as events occurring soon after the initial exposure. As such, future research focusing on how to design a two-phase study in a time-to-event context will likely be motivated by a particular scientific or clinical question and thus be tailored to address issues unique to that analysis.

D. OTHER ANALYTIC APPROACHES

Additional approaches for analyzing two-phase data include approaches drawing on Bayesian methods and propensity score or regression calibration. We provide brief overviews of each approach in the following sections.

1. Bayesian methods

Thus far, we have focused on frequentist methods for analyzing data from two-phase samples. Parametric frequentist methods specify a distribution for observed data that is a function of fixed unknown parameters and then work to estimate these unknown parameters. Using the frequentist estimation approach, all estimation variability is assumed to arise from variability in observed data. Parametric Bayesian methods also specify a distribution for observed data but assume that the unknown parameters are also random and arise from an underlying distribution. Bayesian models specify a *prior distribution* for unknown parameters, and estimation focuses on estimating the *posterior distribution* of parameters given the observed data. The posterior distribution is a function of both the data distribution and the prior distribution. The prior distribution of parameters reflects what is known or expected. In cases where much is known about a scientific problem, prior distributions offer a method for incorporating this knowledge. When little is known, researchers generally specify diffuse or 'noninformative' prior distributions. Often, investigators evaluate the sensitivity of model results to specification of prior distributions, in much the same way as one might evaluate linearity assumptions when including a covariate effect in a linear model.

Bayesian approaches have recently been proposed as an alternative estimation method that is especially useful in situations with sparse data or dependencies, such as data that are correlated.[30] This is particularly relevant to two-phase studies carried out within the context of Mini-Sentinel Surveillance activities, where several factors may contribute to small sample size, including interest in relatively rare outcomes, a desire to detect signals as early as possible (when a relatively small number of outcomes may have accumulated), and time and resource constraints that may limit the size of second-phase samples.

A benefit of a Bayesian approach is that it allows relaxation of the reliance on asymptotic inference, i.e. inference based on the distributions of test statistics and associated statistical significance under the assumption of large sample sizes. Inference based on asymptotic arguments may be inaccurate for small or even moderate two-phase sample sizes of the magnitude expected in Mini-Sentinel activities. Relaxing dependence on asymptotic inference helps to ensure that accurate statements are made about the statistical significance of associations when sample sizes are not large. In addition, Bayesian approaches can incorporate random effects into models to accommodate complex data dependencies (for example, nesting of patients within providers or spatial correlation). Ross and Wakefield (2013)[30] use a log-linear model for the outcome-exposure-confounder relationship, and specify a multivariate normal prior distribution on the set of main effect and interaction terms in the log-linear model. The log-linear model was chosen to accommodate both prospective (as in simple random) and retrospective (as in case-control) sampling. Because we are most often interested in the coefficients from a linear logistic model, it is straightforward to relate the coefficients from the log-linear model to the coefficients from the linear logistic model. Ross and Wakefield (work in progress) have extended this methodology to include random effects terms in the log-linear model to smooth the cell probabilities in large contingency tables, which is particularly useful in sparse data situations. In situations where the exposures and/or confounders take on many values, the resulting contingency table may contain cells with small counts. In this case, through the use of random effects, we can borrow information from neighboring cells to obtain more reliable estimates.

Key Point: Bayesian approaches are an alternative approach to analyzing two-phase data that are potentially useful when sample sizes are small, as may be the case in Mini-Sentinel two-phase studies. Involving a statistician is important when using Bayesian methods to analyze two-phase data to ensure appropriate model specification, assessment of model fit, and estimation.

2. Analyzing two-phase data using imputation approaches

Up to now, we have presented two-phase studies from a survey sampling viewpoint, whereby a phase 2 subcohort is drawn from a larger phase 1 cohort and additional potential confounding factors are measured on this subset. Accordingly, we have described several prominent survey methods to obtain unbiased estimates of an exposure-outcome relationship of interest using these combined phase 1 and 2 data, including re-weighting. Recent applications in epidemiology using such analytic methods include two studies by Breslow et al. involving two-phase stratified sampling in a case-cohort setting. [31,32] Another way to view the data that arise in a two-phase study is from a missing data perspective. In particular, we observe complete outcome, exposure, and confounder data on the phase 2 sample, but the additional phase 2 confounders are missing (by design) for phase 1 cohort members who are not in the phase 2 sample. When viewed from this perspective, it is evident that one could use missing data methods such as imputation to jointly analyze the phase 1 and 2 data. To illustrate, below we

summarize the key steps to implement one missing data approach that has received increasing attention in pharmacoepidemiology study settings: propensity score calibration (PSC).[33]

PSC is a specific application of the regression calibration algorithm described by Carroll et al.[34] It uses a propensity score, the probability of receiving the exposure of interest based on a set of measured variables, to summarize multiple confounders into a single quantity. Phase 1 confounders are used to estimate a crude propensity score for all subjects. Phase 2 confounders are used to estimate a more accurate propensity score among the phase 2 subset. Then, based on the estimated relationship between the crude and the more accurate propensity score within phase 2 sample, a more accurate propensity score is imputed for phase 1 subjects who are not sampled at phase 2. The steps to implement PSC are as follows:

1. Using logistic regression with exposure status as the outcome, estimate an error-prone propensity score (PS_{ep}) for the phase 1 cohort based only on confounders available on all subjects (i.e., the database confounders).
2. Using logistic regression with exposure status as the outcome, estimate a gold-standard score (PS_{gs}) computed in the phase 2 cohort and based both on the database confounders (available for all phase 1 and 2 subjects) *and* on the additional confounders that are only available for phase 2 subjects (i.e., those measured as a result of medical record review).
3. In the phase 2 sample, use linear regression to estimate the association between the predictor PS_{ep} and outcome PS_{gs}, adjusted for exposure status.
4. Use this regression equation to predict PS_{gs} among phase 1 cohort members not in the phase 2 sample.
5. In the phase 1 sample, fit a regression model estimating the association between outcome and exposure, adjusted for PS_{gs} (for those in the phase 2 sample) or the imputed value of PS_{gs} (for those not in the phase 2 sample), and use bootstrapping to estimate the standard errors.

Once data have been gathered for the phase 2 sample, either an imputation method like PSC or a survey sampling method such as reweighting could be used for analyses. The preferred choice will depend on the underlying assumptions one is willing to make, which differ somewhat across these methods. The key element, however, required to reduce bias using either imputation or reweighting is the ability to measure the important confounders in the phase 2 sample. In other words, neither method will perform well if there is substantial unmeasured confounding even after collecting additional phase 2 confounders from medical record review. Both types of methods also rely on the comparability of the phase 1 and 2 samples, which is guaranteed in the context we envision for this methods workgroup since the phase 2 sample is designed as a probability-sampled subcohort. We know exactly what the relationship is between the phase 1 and 2 samples because we defined it. In other instances, we may not understand how the phase 2 subcohort arose. For example, opportunistic data like laboratory variables may be available for some but not all subjects at some Data Partners for reasons that are not transparent.

When PSC is used, unbiased estimation for imputation further depends on the correctness of the model used to impute the gold-standard confounder data from the error-prone information. In contrast, reweighted estimates are robust to this assumption. In other words, they will be no worse than estimates based only on the phase 2 sample, even if this model is incorrect. For both methods, precision will improve as the amount of information in the phase 2 sample increases, which can occur either with

larger phase 2 sample sizes or with increases in the strength of the association between the gold standard and error-prone confounders. Last, imputation requires several further assumptions, including the conditional independence of the error-prone confounders from outcome, given the gold-standard confounders (i.e., the surrogacy assumption).[34,35] That is, once the gold-standard confounders are taken into account, the error-prone confounders are not related to the outcome. Application of imputation using a PS (as described previously for PSC) involves a summary measure rather than a single measured covariate, and this raises additional technical issues many of which have been discussed by Lunt et al.[36] Thus, under certain assumptions, imputation methods like PSC may be useful alternatives to consider when analyzing data from two-phase designs like those that may be used within Mini-Sentinel.

E. "TIME" TWO-PHASE SAMPLING

As part of this work, we also considered the value of ongoing or 'real time' two-phase sampling. This might include ongoing medical record review efforts throughout the course of an initial surveillance activity to validate outcomes or collect more detailed information about confounders. The motivation for such ongoing supplemental data collection might stem from concern that outcome misclassification or confounding could be so strong as to obscure a potential safety signal if one truly existed. This scenario is discussed in more detail in Section IV of this report.

F. SUMMARY AND CONCLUSIONS

In this section, we have reviewed the goals, design and analysis of two-phase studies from a methodological perspective, focusing on application to problems in the context of the Mini-Sentinel program. Two-phase study designs are fairly well developed, with a rich literature in the fields of epidemiology and biostatistics. Much of this work evolved from research examining the performance of case-control studies, and new research is exploring the use of new methods to estimate parameters of interest using data from two-phase designs. As we point out in our next section which focuses on application of two-phase study designs, further methodological work is needed to better guide researchers in both the study design and expected performance of two-phase designs in practical settings.

IV. PRACTICAL ASPECTS OF DESIGNING A TWO-PHASE STUDY FOR SUPPLEMENTAL DATA COLLECTION WITHIN MINI-SENTINEL

A. INTRODUCTION

In this section, we discuss the practical and logistical aspects of designing a two-phase study within Mini-Sentinel. We ground this discussion by working through a hypothetical example based on a current Mini-Sentinel surveillance activity examining the risk of myocardial infarction (MI) associated with saxagliptin use in people with diabetes. We assume that phase 1 data are made up of administrative data from the Mini-Sentinel Common Data Model, and that these are available for a large group of subjects. We also assume that these phase 1 data are used to stratify subjects by exposure and outcome status (and perhaps other factors as well), and then a subgroup of subjects is selected for further data collection in phase 2, most often via medical record review.

We begin by describing the saxagliptin example and our rationale for choosing it. Next, we discuss questions that a study team should ask and answer as they consider designing a two-phase study, and we work through those questions for the saxagliptin example. We briefly discuss a second example, a surveillance activity examining risks associated with use of intravenous immunoglobulin (IVIG), which illustrates the use of a two-phase study design when the major concern is improving measurement of exposure rather than reducing confounding.

B. CHOOSING AN EXAMPLE: RATIONALE AND DESCRIPTION

We began by generating a list of potential examples (Table 7, end of this section). We noted common themes, including special populations (e.g., people with diabetes) and outcomes of broad interest (e.g., cardiovascular outcomes and acute renal or liver injury).

The workgroup chose to focus on the saxagliptin surveillance activity for several reasons:

1. This example is based on a current Mini-Sentinel surveillance activity, so our work might be useful immediately if a signal were to arise.
2. The population (patients with diabetes) and primary outcome (MI) are of broad interest in Mini-Sentinel.
3. The outcome and exposure are well measured in the electronic data. This makes study design considerations simpler than if there were misclassification.
4. Several potential confounders are poorly measured in administrative data *but are likely to be well documented in the medical record*. Thus, this is a realistic example in which supplemental data collection could be helpful.

The protocol for surveillance of saxagliptin has been published,[37] and key features will be summarized here. Saxagliptin is an oral medication to treat diabetes approved in the US in August 2009. In recent years, the recommended first-line therapy for diabetes in the US has been metformin. Saxagliptin is typically used as add-on therapy for patients who are not meeting glycemic control targets despite taking other diabetes medications. Other add-on medications that are therapeutic alternatives to saxagliptin include pioglitazone, insulin, and in some cases sulfonylureas (such as glipizide and glyburide), although the sulfonylureas are still used as an initial medication choice in some patients. The saxagliptin surveillance protocol was undertaken because FDA has a strong interest in assessing the cardiovascular safety of all new antidiabetic medications. This interest stems in part from reports of

excess cardiovascular risk with rosiglitazone use, which were published in 2007. These reports were part of the impetus for updated guidance to industry published by FDA in 2008 requiring that all pre- and post-marketing studies of antidiabetic medications rule out excess cardiovascular risk.

The saxagliptin surveillance study focuses on a cohort of patients with a diagnosis of diabetes who filled prescriptions for an oral antidiabetic drug and had at least 12 months of health plan enrollment. It compares new users of saxagliptin to new users of 4 comparator drugs, conducting 4 separate pairwise comparisons. Comparator drugs are sitagliptin (a previously approved drug from the same class as saxagliptin), pioglitazone, sulfonylureas, and long-acting insulin. In each pairwise analysis, subjects are permitted to add on any other antidiabetic medications except the other drug or drug class in that pairwise comparison. The outcome of interest is acute myocardial infarction (MI). In each pairwise comparison, follow-up ends when a new user stops taking saxagliptin or the comparator drug, has an MI, disenrolls, or reaches the end of the period for which data are available. The analysis is stratified on history of prior cardiovascular disease.

Potential confounders are measured from administrative data and include demographics (age and sex), comorbidity, concurrent medication use, health services use, site or health plan, calendar time, and measures of obesity and smoking (from ICD-9 diagnosis codes). Adjustment strategies include propensity score matching[38-40] and disease risk score stratification.[41-43] Most of the data (including most of the person-years of drug exposure and most of the outcome events) come from Data Partners (DPs) who do not have a unified electronic medical record, and their electronic data do not include vital signs or lab values. A small proportion of the exposed person-time comes from integrated healthcare delivery systems with electronic data available for vital signs (e.g., blood pressure, height, weight) and lab values (e.g., cholesterol level and hemoglobin A1c) measured as part of routine clinical care.

There are several potentially important confounders that may obscure the relationship between drug exposure and the outcome (MI) in the saxagliptin study, including diabetes severity, smoking status, obesity/body mass index, and levels of blood pressure and lipids. These are either unavailable or poorly-measured in administrative data. Thus, if a signal were to arise, the team would consider reviewing medical records to obtain better measures of these potential confounders.

C. QUESTIONS TO BE ASKED WHEN DESIGNING A TWO-PHASE STUDY

Table 8 (at the end of this section) lists questions to consider in designing a two-phase study. We now discuss each of these questions and consider them in the context of our example, surveillance of saxagliptin.

1. What are potential key confounders?

General considerations: To be a confounder, a characteristic must be associated both with the exposure of interest (here, use of a medical product) and the outcome. Thus, known risk factors for the outcome should be considered. If a risk factor is associated with the exposure (because it either influences the exposure or shares a common unmeasured cause with the exposure), then it should be included as a potential confounder. It is important to understand that the outcome risk factors could influence receipt of the product through mechanisms that are indirect. For instance, a risk factor that increases the frequency of clinical visits or laboratory tests increases contact between patient and provider and through this mechanism may increase the likelihood of a new medication being initiated.

Confounding by indication is an important consideration in pharmacoepidemiologic studies. This type of confounding arises when a drug is prescribed for an indication that is associated with the risk of the outcome of interest.[44] A classic example is the case of beta blockers and acute myocardial infarction. Beta blockers are used for several indications including angina, a symptom of coronary artery disease. Because of confounding by indication, beta-blockers could appear to increase risk of myocardial infarction compared to other antihypertensive medications that are not also used to treat angina. In fact, the apparent increased risk arises because beta-blockers are often prescribed to patients who are at high risk for myocardial infarction due to existing coronary artery disease. Thus, the indication for use of the drug or device under study should be considered. Severity of the indication is also a potential confounder, since patients with a more severe (or poorly controlled) health condition may be more likely to initiate a new treatment. In some Mini-Sentinel activities, restriction may be used to address confounding by indication: the study population may be limited to people with the particular health condition for which the drug is typically prescribed. Even in this case, concern may remain about severity of the underlying condition, which could vary between patients who initiate the new medication and those who do not initiate use.

Markers of overall poor health status (or alternatively, robust health status) are another important type of confounder. These markers include measures of poor functional status, such as the need for assistance walking or bathing, as well as nursing home residence and the need for equipment such as wheelchairs or home oxygen therapy. Prior studies have demonstrated that such health status measures can be important confounders, for example in studies of mortality in relation to influenza vaccine[45,46] and outcomes of community-acquired pneumonia in relation to statin use.[47] Health status confounders are of particular interest in the context of a two-phase design because they are typically not available in large electronic (administrative) datasets. These measures may also be difficult to obtain from review of medical records, although several studies have demonstrated that adjusting for imperfect measures of functional status obtained from medical records improves control of confounding.[45,46]

Research studies often seek to measure overall health status via comorbidity indices calculated from administrative claims data, such as the Charlson score or others. There is evidence, however, that such scores may not adequately capture health status, particularly in older individuals. One example comes from a study of mortality in relation to use of a wide variety of medication classes where paradoxical relationships were observed, such as lower mortality in patients receiving nonsteroidal anti-inflammatory drugs or hypnotic medications compared to nonusers.[48] These associations likely reflect clinicians withholding these medications from patients who are frail or nearing death. Adjustment for the Charlson index did not alter the results. A second example comes from a study of mortality in relation to influenza vaccine,[45] a setting where strong "healthy user" bias has been shown to be present.[45,49] Adjusting for components of the Charlson score (a variety of comorbid illnesses) measured from administrative data did not reduce bias but in fact moved risk estimates further from the null. In contrast, adjusting for measures of functional and cognitive status obtained through medical record review moved results toward the null, suggesting better adjustment for "healthy user" bias. These studies and others have suggest that comorbidity indices measured from administrative claims data have important limitations in terms of their ability to measure overall health status in older adults.

Characteristics that influence early adoption of a new medication or device are another type of potential confounder, especially when the new treatment modality is being compared to existing alternatives. Patients with severe or difficult-to-control disease may be more likely to adopt a new therapy soon after

its introduction, because the patient and/or clinician hope that the new treatment will succeed where prior therapies have failed. Early adoption of a new therapy may also be more common in younger patients and those with higher socioeconomic status,[50] potentially leading to healthy user bias. Characteristics of the physician and practice settings influence early adoption; for instance, physician specialty, younger physician age, the availability of facilities for specific procedures, and practicing in an academic center are associated with early adoption of new therapies.[51,52] If these characteristics are also associated with risk of the outcome of interest, then they should be considered as potential confounders. The characteristics that influence early adoption are likely to be specific to each drug or device and to the medical specialty involved. This implies that surveillance activities will benefit from input from practitioners and exploration of available data to better understand this type of potential confounder.

Application to saxagliptin example:

Table 9 at the end of this section shows the confounder list generated by the two-phase studies workgroup based upon consideration of the questions outlined above. The list includes disease risk factors (smoking, obesity, blood pressure, lipid levels); measures of diabetes severity such as duration of diabetes and level of hemoglobin A1c; and measures of poor health status (nursing home residence).

2. Are potential confounders available in the phase 1 (electronic) data?

General considerations: If good-quality measures of potential confounders are available in the phase 1 (electronic) data, then it is important to consider whether previous analyses have fully utilized these data. Increasingly, future Mini-Sentinel surveillance activities will use standard modules that draw on a general list of confounders. When a signal first arises, some relevant confounders that are in fact available in phase 1 data may not yet have been included in analyses. Before proceeding to a two-phase study, analyses should be carried out that take advantage of the full range of confounder measures available in phase 1 administrative data.

If information about potential confounders is available in the phase 1 data, then the next step is to consider the accuracy of these measures. Some illnesses and procedures are very well measured in administrative data. For example, cardiac procedures are very well measured. One recent study reported that the positive predictive value (PPV) for administrative-data measures of coronary artery bypass graft surgery was 96% and for percutaneous coronary intervention, 94%, compared to the gold standard of a clinical registry.[53] Other conditions are known to be less well measured; for instance, the sensitivity of administrative-data measures was 76% for congestive heart failure and 61% for hypertension in one study.[54] Diagnostic codes are particularly problematic for measuring behavioral risk factors: for instance, the sensitivity of administrative-data measures for smoking was estimated to be only 7% in one study,[55] and the sensitivity of ICD-9 codes for obesity in children was estimated to be just 15%.[56] In the most extreme case, there may be no capture of important potential confounders in phase 1 data (that is, the Mini-Sentinel Common Data Model datasets). In this case, it will only be possible to measure and adjust for confounders through supplemental data collection. Two-phase studies can also be helpful in the less extreme situation where some information about confounders is available in the phase 1 data but the available measures are inaccurate.

The accuracy of confounder measurement in the phase 1 data may be affected by the length of the baseline period chosen in a specific surveillance activity. The “baseline period” refers to the period of prior (or “baseline”) health plan enrollment required before patients are included in cohorts used for

surveillance activities. The baseline period often ranges from 6 to 12 months, and it is generally used for 3 purposes: 1) to measure baseline patient-level covariates, prior to exposure; 2) to identify a cohort of "new users" of the medications of interest by requiring a 6 to 12 month period with no exposure to these medications; and 3) to ensure that outcomes observed during follow-up are truly incident. Choosing a relatively short baseline period may lead to inaccurate measurement of some confounders, particularly those that relate to historical events, e.g., a history of MI or coronary artery bypass grafting (CABG). Measurement challenges are more likely to arise when remote health events as well as recent events are of interest. It is important to note that even when the events of interest are well documented in administrative data at the time they occur (e.g., MI or CABG), ascertainment of historical events may be incomplete if the event occurred prior to the designated baseline period or the patient was not enrolled in the health plan at the time. One implication is that it is important to consider the impact on confounder measurement when selecting a baseline period for surveillance activities.

There may be cases where a potential confounder of interest is not well measured in administrative data, but a proxy measure is available that is believed to be more accurate. For example, to measure hyperlipidemia, it may be preferable to use pharmacy data to identify receipt of a statin, rather than seeking diagnosis codes for hyperlipidemia. Given the vagaries of medical coding, we may suspect that there is undercoding of hyperlipidemia, so a variable based on diagnosis codes would lack sensitivity. Since statins are given for a limited number of indications, we can be fairly confident that people receiving a statin have hyperlipidemia. Of note, this proxy measure does not provide an indication of the severity of hyperlipidemia, which is difficult to assess other than through laboratory data.

When proxy measures are available in phase 1 data, investigators should also consider whether additional information that can only be obtained through supplemental data collection would substantially improve control of confounding. For example, it may be desirable to obtain measures of blood pressure values or lipid levels to improve control of confounding, instead of using a binary variable for the presence of hypertension or hyperlipidemia.

Application to saxagliptin example:

Table 10 (end of this section) summarizes the workgroup's thinking about the accuracy of phase 1 data for measuring potential confounders relevant to saxagliptin surveillance. We did not conduct a systematic literature review for this illustrative example but rather drew on workgroup members' knowledge and past experience in studying these conditions. We also drew on other expertise available within Mini-Sentinel and where possible obtained data about validity of these measures from the literature. To summarize: in workgroup members' experience, the presence or absence of some conditions (e.g., hypertension, hyperlipidemia) may be reasonably well measured in administrative data, but it is challenging to measure duration and severity. It would be desirable to adjust for blood pressure values and cholesterol levels as severity measures. However, these data are not available within electronic claims data, the kind of data provided by the Data Partners contributing the largest amount of exposed person-time and presumably the majority of outcomes in the saxagliptin study. For some conditions (smoking and obesity), phase 1 measures are expected to have very low accuracy. Administrative data should be able to supply good to excellent measures of nursing home residence but will lack other indicators of frail health. For instance they do not routinely or reliably capture functional or cognitive impairment.

3. Can confounders be accurately measured from other sources (e.g., medical records)?

The primary purpose of two-phase sampling is to measure important confounders that are unavailable or poorly captured in administrative (phase 1) data. Therefore, availability of more accurate confounder information is essential for a successful two-phase study. In theory, there is a range of possible methods for two-phase data collection, including medical record review, mailed questionnaires, telephone interviews, collection of biological specimens, and in-person examination and interview. In practice, medical record review is most likely to be used, for a number of reasons. From a scientific perspective, the use of medical record review minimizes biases including recall bias, which can arise when patients are interviewed after the outcome has occurred (there can be differential recall in patients with and without the outcome of interest.) When data are collected from patients, selection bias can arise because it is not possible to collect new data from patients who have died. Moreover, medical record review is less burdensome for patients than other approaches, and in general it is more convenient and less costly than approaches that involve contacting patients.

Documentation of confounders in the medical record will vary greatly and will depend on the context, including the population of interest and the time period that is chosen as the focus of chart review. When the population of interest has a chronic illness such as diabetes that requires regular office visits, then documentation is expected to be more complete. Laboratory and other diagnostic tests may be carried out more frequently for people with chronic illnesses, compared to people without such illnesses. Similarly, confounder information may be better captured in the medical record for people who are "new users" of medications because these patients are, by definition, engaged with medical care. Completeness and accuracy of documentation may also be better for information that is relevant to national quality measures or quality improvement initiatives. For instance, for patients with diabetes, quality measures include measuring hemoglobin A1c at certain time intervals and achieving A1c levels below certain thresholds. Health plans and providers often have strong financial incentives to comply with these measures. Thus we expect to find more information in medical records about these kinds of test results than for laboratory tests that are not relevant to any quality measures. In addition, some health care systems have processes in place to routinely collect and record some characteristics in the medical record. For instance, in some systems, smoking status is considered a "vital sign" and is queried and documented by clinic staff at each office visit. This process may rarely result in an ICD-9 diagnosis code, and as a result smoking status will be poorly measured in administrative (billing) data, yet it may nonetheless be readily available in the medical records themselves. Diagnoses of chronic illnesses that require ongoing care (e.g., chronic heart or lung disease) are likely to be relatively well recorded in medical records. Similarly, invasive procedures or surgeries are likely to be well recorded (e.g., CABG), though as discussed above, the documentation may have occurred at the time of the event and not be captured in the shorter baseline period chosen for a given surveillance activity.

Some confounders may be poorly documented in medical records. For example, information in the medical record may be less complete for conditions that carry stigma (e.g., alcoholism, substance abuse, mental illness), in part because patients may be reluctant to seek care. Cognitive and functional status, important potential confounders in studies of medication safety in older people,[45-47,57] may also not be well documented in medical records. Reasons for incomplete documentation include that some patients with dementia do not present for evaluation until their symptoms are severe. In other cases, clinicians fail to recognize or document that dementia is present. Functional status (such as ability to ambulate independently or perform other activities of daily living) may not be routinely documented. However, there is some reason to think that documentation of cognitive and functional status may be improving.

New regulations related to Medicare billing are promoting more frequent assessment and documentation of functional and cognitive status as part of Annual Wellness Visits, and many health care systems have implemented standard templates and processes to ensure better documentation. These kinds of process changes will likely increase the availability of information about functional and cognitive status in the medical record. This would be a boon for epidemiologic studies which seek to measure these characteristics through medical record review.

Related to the problem of missing information is the problem of selective documentation, which can lead to bias. One concern is that better documentation may be available for patients who are seen more frequently in a given health care setting. In patients who rarely seek care, documentation may be scant, and information on key confounders may be missing. The underlying mechanisms that drive differences in the frequency of healthcare visits are not well understood and probably vary across individuals. For instance, some patients may rarely seek care because they are extremely healthy (no chronic illnesses and only rare injuries or acute illnesses), while others may be gravely ill and have difficulty with mobility and transportation. Concerns about selective documentation may be reduced by a new user design that relies on an active comparator. In this setting, the fact that exposed and comparator groups have both started a new medication recently is likely to ensure more comparable documentation than if exposed people were compared to people not starting a new medication. Concerns may be reduced further if it is possible to restrict to patients who are all at the same stage of their illness, for example initiating treatment for their diabetes for the first time.

The choice of a time period to focus on has obvious implications for data availability. This design choice is discussed in more detail below (Question 10.) Briefly, a team planning a two-phase study must specify the time period during which information will be sought in the chart. Often this will be a period of time immediately before initiation of a drug of interest. Choosing a relatively short period of time (e.g., 3 or 6 months) increases the risk of missing data, particularly if patients with the underlying health condition are not seen very frequently in clinic. Certain labs may be measured at regular intervals that are longer than 3 months (e.g., fasting lipids once a year, hemoglobin A1c every 6 months.) Choosing a longer time window, e.g., 2 years, may decrease missing data but will require more time and resources.

Application to saxagliptin example:

Several aspects of the saxagliptin example increase the likelihood that confounder information will be recorded in the medical record. First, the population is patients with diabetes, who are more likely to have frequent clinic visits and regular laboratory monitoring than people without a chronic illness. Second, for diabetes, there are established quality metrics that promote regular monitoring (e.g., testing hemoglobin A1c and cholesterol levels). In addition, this surveillance activity focuses on patients initiating a new medication – either saxagliptin or an active comparator. This medication initiation is more likely to occur when a health care provider notes that a patient's glycemic control is poor – which is likely to coincide with a clinic visit and/or laboratory test result. For all of these reasons, we anticipate that patients in the saxagliptin study will have abundant documentation in the chart.

The accuracy of measures in the medical record is likely to vary across confounders. Many measures of interest – such as laboratory values (hemoglobin A1c, cholesterol) -- are likely to be highly accurate. We expect weight to be routinely measured at each office visit, though height may be missing for some patients, resulting in missing BMI. Blood pressure measures are likely to be available almost universally, although recommended processes for measuring blood pressure may not always be followed in busy

clinical practices. We expect smoking status to be quite complete, given the strong emphasis on reducing cardiovascular disease risk in patients with diabetes, although it may not be assessed at every visit. Availability of smoking information may also vary across charts, for instance by care setting, health care system or provider specialty.

Diabetic complications (such as retinopathy and nephropathy) were included as potential confounders because they can shed light on certain aspects of diabetes severity. We expect that in general, providers will routinely assess and document these at diabetes-focused visits, probably at least annually. Some patients do not come in routinely for recommended care, and some providers may fail to document these measures, but we suspect that in the patients in the saxagliptin study cohorts—all of whom are making a medication change--completeness of these measures will be relatively high.

Comorbid illnesses such as congestive heart failure, history of coronary artery disease, and cancer (by which we mean recent or active disease) are expected to be well documented. As discussed above, measuring history of coronary artery disease (including procedures such as CABG and percutaneous coronary intervention [PCI]) is challenging because accurate measurement may require evaluation of information over a long time period. Historical events such as MI may be documented at the time of their occurrence or at a patient's first visit, but a patient's prior medical history may not be consistently or thoroughly documented at follow-up visits. Thus, if chart review is limited to a 6 or 12 month period prior to initiation of a new medication, there is likely to be some misclassification of patients' history of coronary artery disease (CAD). A more detailed discussion of considerations in selecting the time period for chart review is found below under Question 10. Typically, a medical record includes a "problem list" section intended to summarize all major health conditions; this list is supposed to be maintained and updated over time. Thus, one solution to improve the capture of past major health events is to utilize the "problem list" as a source of information. Overall, we expect the measures of CAD available from the chart to be better than those routinely available in administrative data covering the same time period.

We considered including measures of functional and cognitive status as potential confounders of interest. Ultimately, we elected not to include these measures because of concern that these are relatively poorly documented in medical records. Also, they could be selectively documented, leading to bias. We did include nursing home residence as a proxy for poor health status. We expect documentation of this characteristic to be quite complete in the medical record because nursing homes need physician orders and review and approval of care plans on a regular basis. However, since claims data may capture nursing home care quite well (see Table 10), ascertaining this characteristic through medical record review may not improve much on measures obtained from the Phase 1 data.

4. Should the phase 2 sampling scheme stratify on and oversample any confounders?

As outlined in Section III of this report, it is generally desirable to select the phase 2 sample from strata created by cross-tabulating exposure and outcome status. This approach enables oversampling of people with both the exposure and the outcome, likely a small but highly informative group. When there is a strong confounder that is also rare, and a proxy measure for that confounder is available in phase 1 data, it could in theory be helpful also to oversample patients with the (phase 1) proxy for the (phase 2) confounder. Such oversampling will greatly increase the prevalence of the rare confounder in phase 2, potentially enabling better adjustment for that confounder. In contrast, if patients are sampled without regard to a rare confounder, then the phase 2 sample may contain so few patients with this characteristic that adjustment for the confounder is either inadequate or impossible. There is no specific

guidance available about what cut-off should be used to define confounders as "rare." The decision about whether to oversample people based on their confounder proxy status should also take into account the size of the phase 2 sample. If the phase 2 sample is relatively small, then further stratification by confounder status may not be wise, as it may increase the number of very small or even empty cells.

The exact setting in which there would be benefit from oversampling on a phase 1 proxy for a rare confounder is unknown. Whether there would be efficiency gains (and how much) will depend on characteristics of the confounder such as its prevalence and the strength of its relationship with the observed data, as well as the quality of the surrogate measure. Additional methodological research is needed to untangle the relative importance of these factors with respect to efficiency. In the context of Mini-Sentinel surveillance, guidance about whether to stratify on a particular confounder surrogate may come from simulations. A simulation tailored to a specific scenario could shed light on whether it is likely to be helpful in that instance. Later on, in Section V of this report, we discuss the use of simulation studies to help guide decisions about study design for two-phase studies within Mini-Sentinel.

Application to the saxagliptin example:
Most potential confounders in the saxagliptin example (e.g., hypertension, hyperlipidemia, obesity) are expected to be relatively common. However, three confounders are expected to be relatively rare: current smoking (perhaps 10-15%), residence in a nursing home, and a current cancer diagnosis (meaning a recent cancer diagnosis or ongoing cancer treatment, rather than a history of cancer.) Of these confounders, smoking is likely to be measured most poorly from the automated data. Thus, when selecting the phase 2 sample, the team should consider whether to oversample people who are likely to be smokers based on proxy variables for smoking status available at phase 1. These proxy variables include ICD-9 diagnosis codes for tobacco abuse. In theory, such oversampling might be helpful, but in fact there is little clear guidance from the literature. Our own limited simulations (presented in Section V of this report) suggest that in this particular case stratifying on smoking status would not be beneficial. The simulation results are influenced by the specific parameters we chose for the prevalence and associations of the confounder, exposure and outcomes. So if serious consideration were being given to stratifying the population on smoking status, we would want to conduct a more extensive simulation considering a range of values for these parameters.

5. Does the outcome require validation?

Supplemental data collection may offer an opportunity to validate outcome status, which is desirable when the outcome of interest is poorly measured in administrative data. When electronic data algorithms have low positive predictive value (PPV) for an outcome, the resulting misclassification can cause substantial bias. In a recent study, administrative data codes for rhabdomyolysis had a positive predictive value of only 7.5%.[58] The authors estimated that relying on administrative data in a study of statin use and rhabdomyolysis would have severely biased results toward the null; a true incidence rate ratio of 2.6 (95 % CI 1.0-7.8) would have been attenuated to 1.0 (95% CI 0.8-1.3) had the analysis defined outcomes solely from ICD9 codes. Other examples include severe acute liver injury, which in a Mini-Sentinel validation study had PPV 25% in people without and 41% in people with chronic liver disease.[11]

Whether it is worthwhile to attempt to validate outcomes at the same time as measuring confounders will depend on the situation. Some Data Partners (DPs) have integrated electronic medical records,

making it relatively efficient to collect information about confounders and outcomes at the same time. At other DPs, patients' care is dispersed across a variety of settings including inpatient, outpatient, primary and specialty care, with each setting maintaining a separate chart. At these DPs, validation of outcomes may require a second, independent chart review, which is inefficient. It will be important to decide whether there is greater concern about bias due to residual confounding or outcome misclassification. Available resources should be focused on reducing the bias that is felt to pose the greatest threat to the validity of results from surveillance activities.

Application to the saxagliptin example:

The primary outcome of interest in the saxagliptin example is acute MI, which is well measured in administrative data. A validation study within Mini-Sentinel reported a positive predictive value of 86%.[10] Thus, supplemental data collection should focus on confounder information.

6. Are there aspects of exposure that need to be measured from medical records?

In some scenarios, important aspects of exposure may be difficult or impossible to measure from administrative data. For example, for the biologic product intravenous immunoglobulin (IVIG), the brand of the product as well as concentration, dose, and rate of administration may affect the risk of adverse outcomes. IVIG is used to treat a wide range of medical conditions, and the underlying risk of various outcomes may also differ according to the medical condition that IVIG is being used to treat. These types of information may be absent from or poorly captured in administrative data but may be documented in the medical record. Again, whether it is easy to collect information about exposure and confounder status in the same chart review will depend on the context and the DPs. At first glance, it seems that it would be more likely that a single chart review could collect both confounder and exposure data (in contrast to outcome data), because the confounder information may be recorded at the clinic visit where the product is ordered or administered. Details regarding selection of which DPs and charts to pursue for supplemental data collection will be discussed in more detail in Questions 8 through 10 below.

There is another aspect of exposure status that can be difficult to assess from administrative data: prior history of exposure to the medications of interest. Mini-Sentinel activities often utilize a "new user" design. One rationale is that "prevalent user" cohorts include a large proportion of people who have already tolerated a medication well without adverse effects. Susceptible individuals may have experienced an adverse effect soon after initiation and thus be underrepresented due to a phenomenon called "depletion of susceptibles". To operationalize the "new user" design, Mini-Sentinel activities define "new users" as people with no fills for a study medication in the baseline period, say 6-12 months. This approach fails to identify people who used the medication in the more distant past, who again presumably survived exposure without serious adverse effects. Thus there may be misclassification of "new user" status. While such misclassification may be rare for newly introduced medications (because surveillance begins soon after a drug comes on the market), this misclassification may be more common for comparator drugs, some of which may have been on the market for many years. Including former users who did well on the active comparator in the past (one possible reason for restarting it) could enrich the comparison group for people unlikely to develop adverse effects. This could lead to an apparent increased risk of adverse effects in the group exposed to the primary medication of interest. Since it may be difficult to identify true "new users" from administrative data, medical record review could be used to confirm "new user" status for all study drugs. This approach will be most useful if relevant information about prior medication use is expected to be routinely recorded

in the chart when a medication is being initiated. It will be less feasible, and pose additional challenges, if abstractors must search the entire record for this information. Since different people may have different lengths of prior enrollment, there is potential for differential measurement error, which could lead to bias, as is discussed further in Section IV. C. 10 (below). To summarize, it is unlikely that the desire to confirm "new user" status would be a primary motivation for medical record review, but if reviews are being conducted for other reasons, it is worth considering whether incorporating questions about this topic might be useful.

Application to the saxagliptin example:

In the saxagliptin example, prescription fills (including refills and numbers of "days supplied") are well measured by electronic data. It would also be helpful to ascertain whether patients actually took their medications as prescribed, but medical record review is unlikely to improve measurement of this. Administrative data may misclassify "new user" status for comparator drugs that have been marketed for many years, e.g., sulfonylureas. The team might consider gathering information about past use during medical record review. However, it will be important to evaluate how much time is required to collect this information, especially if data collection requires reviewing past chart notes far back in time. This information will be most feasible to collect if reviewers find that past use is mentioned around the time of initiation of a medication.

7. Do records need to be reviewed to clarify the timing of the outcome in relation to the exposure?

In some scenarios, it may be important to clarify the timing of the outcome in relation to the exposure. One such situation is when an outcome occurs very soon after medication initiation. Examples include anaphylaxis occurring soon after exposure or an acute thrombotic event (MI or stroke) occurring during or soon after infusion of intravenous immunoglobulin (IVIG.) In these cases, it may be particularly important to determine whether the exposure truly occurred prior to the outcome. Another situation where the relative timing is critical is when the analysis specifies certain "risk windows" in which exposure is hypothesized to cause the outcome. An example would be a self-controlled case series design which compares risk of the outcome in one time window to risk in a different time window. In this setting, knowing the exact date and time of exposure may be needed in order to assign outcomes to the correct "risk window."

The electronic data available in Mini-Sentinel have limitations that can make it difficult to determine precisely when exposures and outcomes occurred. Pharmacy dispensing data show when a medication was dispensed, but the medication may in fact be consumed or administered on a future date. Diagnosis data indicate the date (or date range, for an inpatient stay) when healthcare was provided for a given diagnosis, but they do not pinpoint when the condition first developed nor, for inpatient stays, on which day(s) of the hospital stay a diagnosis was present. For some health conditions, patients may have symptoms for some time before seeking care. Other health conditions may develop during the course of a hospital stay, e.g., nosocomial infections. Thus, when exposure and outcome are documented on the same day or during the same hospital stay, the temporal sequence often cannot be determined from the automated data. Challenges also arise when studying medications or products used for people with severe illnesses that lead to prolonged hospitalization, as these individuals are at higher risk for many adverse outcomes (e.g., venous thromboembolism; gastrointestinal bleeding) during hospitalization due to their severe illness. In these situations, medical record review can be helpful to clarify the temporal

relationships between outcome and exposures. Two-phase study designs may be useful to help target medical record review to the most informative individuals.

Application to the saxagliptin example:
In the saxagliptin example, there is little concern about the timing of the outcome in relation to the exposure. Saxagliptin is expected to be initiated in the outpatient setting, not during a prolonged hospital stay. It is extremely unlikely that a patient would experience an acute MI on the day that s/he fills a saxagliptin prescription. Also, MI is a severe, acute event; in general, we do not expect that a patient would have gradual onset of symptoms or prolonged symptoms prior to seeking care and receiving a diagnosis of MI.

8. At which Data Partners should phase 2 data collection be carried out?

The Mini-Sentinel program currently includes 18 Data Partners (DPs). The goal in carrying out a two-phase study is to collect supplemental data as efficiently as possible. It is not desirable or feasible to spread medical record reviews equally over all 18 DPs, either from a logistical or scientific perspective, and so some small subset of DPs must be selected for more detailed data collection. Logistically, it will be more efficient (and less time-consuming) to limit chart review to a relatively small subset of DPs, so that there can be efficiencies when requesting, reviewing and redacting charts. There are also scientific reasons to minimize the number of participating DPs. As we discussed in Section III, the phase 2 sample will typically be selected from strata that are created by cross-tabulating exposure versus outcome status, yielding 4 cells. Unless investigators believe that patients are exchangeable across DPs – that is, that patient characteristics, physician practice styles, and data capture are similar - this number must be then multiplied by the number of DPs. If 4 DPs participate in chart reviews, this will yield a minimum of 16 strata. With so many strata, the likelihood of small cell counts increases which can decrease precision and bias results from a two-phase analysis. Section V of this report, which focuses on simulation studies, highlights the problem of having overly small strata. Thus, chart review should be carried out at a relatively small subset of DPs, perhaps 2 to 4, depending on the number of medical records that the team plans to review.

Existing literature about two-phase studies does not provide guidance to direct the choice of DPs in this setting. This problem is specific to the distributed data environment of Mini-Sentinel. While development of guidelines for selecting DPs would be valuable for future Mini-Sentinel projects, such research is beyond the scope of this workgroup. Below, we provide some initial thoughts about DP selection and the rationale for this guidance.

First, the team should consider at which DPs the signal was observed. DPs where the signal emerged should be a high priority for supplemental data collection. The primary focus of supplemental data collection is to investigate whether a potential safety signal is driven by patient selection and confounding. If the results of a two-phase study show that the signal at these DPs was most likely due to confounding, then this should greatly diminish concerns about the initial signal. It may also be of interest to include one or two DPs where no signal arose, particularly when the direction of confounding is uncertain—that is, when we are unsure whether confounding caused a spurious signal to arise at certain DPs or a true signal to be obscured at others.

A second consideration is the distribution of exposed person-time and outcomes across DPs. If some DPs have contributed very little exposed person-time, or few or no outcomes, then it likely will not be helpful to include them in phase 2 data collection.

A third consideration is whether certain sites are more likely to have information available about relevant confounders that are the focus of phase 2. For instance, certain DPs have comprehensive electronic data on laboratory results and vital signs covering all or nearly all members. In the most extreme case, if a DP is not expected to have additional confounder information available in the medical record, then including that DP in the second phase will not provide any useful information.

A final consideration is the efficiency of chart review. At integrated health care systems, all information (inpatient and outpatient, primary and specialty care) resides in a single unified electronic medical record. Focusing on such settings will improve efficiency compared to obtaining multiple charts (or a single chart with incomplete information) from a DP where care is more dispersed. On the other hand, in some Mini-Sentinel activities, the DPs that are integrated healthcare systems may contribute considerably less exposed person-time and fewer outcomes compared to the largest DPs, which generally are those where care is most dispersed. So the desire for efficiency may come into conflict with the second consideration above, the need to carry out medical record review in the settings which contribute substantial proportions of exposed person-time and outcomes.

Selecting only a few DPs for phase 2 data collection has potential to introduce bias, because their populations (and therefore the samples selected from them for review) may not be representative of the entire (target) population. The 18 DPs are heterogeneous in many respects, including patient population (e.g., age, racial/ethnic distribution) and geographic region. There may be considerable differences in practice patterns (including early adoption of new therapies) across DPs, some due to policies (e.g., formularies) and some related to the clinical culture of a particular health care system or geographic region. Some DPs represent integrated health care systems, while others are much more decentralized, spanning a larger geographic area and including a large number of physician groups and health care systems. Other DPs are a hybrid. It is also likely that results of initial analyses may be heterogeneous across the DPs. It is important to recognize that by selecting a limited number of DPs in phase 2, the charts being reviewed can never be truly “representative” of the entire study population. At the same time, this sampling is inevitable, given constraints on the phase 2 sample size. If failure to adequately represent all DPs is of particular concern, the team may wish to target DPs that represent different characteristics for which there is heterogeneity, such as DPs where a signal did occur and where it did not, as well as DPs that represent integrated healthcare systems and those that do not.

In summary, data collection in phase 2 must target only a few DPs. We recommend focusing primarily on 1) DPs where a signal is seen and 2) DPs contributing a relatively large amount of exposed person-time and outcomes. Additional important considerations are the likely availability of confounder information, efficiency of review, and a desire to maintain some balance in the DPs selected, based on recognizing the heterogeneity of the underlying study population.

Application to the saxagliptin example:

Because no signal has yet arisen, we cannot determine which DPs should be the focus of phase 2 data collection. However, data suggest that the great majority of exposed person-time and outcomes came

from a few large DPs, so we anticipate that one or more of these DPs would be chosen for supplemental data collection if a signal were to arise.

9. If medical records are to be reviewed, from which care setting and/or provider should these be obtained?

At some DPs, all information is consolidated in a single unified electronic medical record. When reviewing records at these DPs, this question is not applicable, as all information can be found in one place. However, at other DPs, care is much more dispersed. A patient may receive care in multiple settings (inpatient, outpatient, emergency department) across multiple health systems, all of which maintain separate medical records. Even within a single health care system, inpatient records may be contained in one chart while outpatient records are maintained separately. When reviewing records from these DPs, the team will have to specify which medical record is to be reviewed. The team might consider obtaining and reviewing multiple records for each patient, but this will greatly increase the resources and time required, and may require overall sample size to be smaller, which poses other challenges (further explored in Section V of this report.)

When reviewing records from a DP where care is more dispersed, the choice of which record to review depends on the exposure, outcome, and confounders being considered. When a surveillance activity is using a "new user" design, then a reasonable approach is to focus on the setting where the medication of interest was first prescribed. The rationale is that the provider who wrote that prescription relied on the information available in that medical record when deciding to prescribe the medication of interest. Thus, the confounders most likely to have shaped the decision to prescribe that particular medication are more likely to be identified in that medical record than in records from other care settings.

Application to the saxagliptin example:
We expect that in most cases, saxagliptin is initiated in the outpatient setting by the provider managing the patient's diabetes. This may be a primary care provider, endocrinologist or cardiologist. Thus we recommend obtaining the medical record from the prescriber who first prescribed saxagliptin. In a few cases, saxagliptin might be initiated during an inpatient stay when a provider notes that a patient's diabetes is poorly controlled. In this case it would be reasonable to review that inpatient record.

10. What time period should phase 2 data collection target?

The choice of a time period for phase 2 data collection must balance competing concerns including concerns about 1) missing data; 2) selection bias; 3) differential measurement error; and 4) resource constraints. If supplemental data collection focuses on a relatively recent time period, say the 6 months before a new medication is initiated, then there is a considerable risk that some important confounder data will be missing. Not all people are seen in clinic every six months. Some labs, such as fasting lipid levels, may be measured only annually or less often. This shorter time window may not include important historical events, such as MI or CABG. On the other hand, one could argue that characteristics or conditions not mentioned in the medical record for 6 months or more are not likely to influence prescribing decisions, reducing the likelihood that they will be strong confounders.
Selecting a longer time period (e.g., 5 years) raises other concerns. If this decision is accompanied by a decision to require longer baseline health plan enrollment, this decision changes the population being studied. It may dramatically decrease sample size for the overall analysis, thus reducing power. It may also change the characteristics of the population, because people with long and stable enrollment in a

given health plan may differ from people who switch plans. These differences may be difficult or impossible to measure. As a result of this enrollment requirement, the population being studied may no longer be representative of the general population whose risk of adverse events we wish to understand. A third option, even more problematic, would be to leave enrollment requirements the same but look within medical records over whatever time period is available. In this approach, information is collected over different durations of time for different people, which may lead to differential measurement error. People with a longer duration of enrollment will have less missing data than people with a shorter duration of enrollment, and again, people with stable, long-term enrollment may differ in important ways from people who switch plans. If these differences are related to the risk of adverse events, then bias may result, and the direction and magnitude are difficult to predict.

Logistical considerations also affect this decision. Reviewing medical records over a longer time period will take substantially more resources than focusing on a shorter time period. This may limit the number of records that can be reviewed in phase 2, which can affect the overall usefulness of supplemental data collection, as is discussed in greater detail in Section V of this report.

On balance, considering all of the factors discussed above, the two-phase study design workgroup felt that in many cases, 12 months would be a reasonable time window for the collection of supplemental data, in conjunction with use of the active problem list from the medical record. For activities focusing on medications administered for acute conditions in the inpatient setting, a much shorter time period may be acceptable. In any case, the time window will need to be tailored to the specific needs of a given surveillance activity.

A separate but related question is during what time period relative to exposure the confounder information should be sought. It seems clear that confounders should be measured during the period prior to medication initiation. But might there also be value in some cases in seeking confounder information after baseline, that is, during the follow-up period? For example, if initial phase 1 analyses reveal substantial differences across the study drugs in regard to medication adherence, and if the phase 1 analyses censor people when they discontinue medication use, then further investigation of the reasons for cessation may be helpful. One motivation is to allay concerns about bias due to informative censoring. By informative censoring, we mean that people who stop the drug (and thus drop out of the analysis) may differ from those who continue (and remain in the analysis) in terms of their risk of the adverse outcome under study. Further discussion of this situation is outside the scope of this report, but it provides another example of a scenario where supplemental data collection via a two-phase study design may be useful.

Application to the saxagliptin example:

The workgroup felt it would be best to review medical records covering the 12 months prior to initiation of saxagliptin or the active comparator drug. They felt that the benefits of going back farther in time would be outweighed by concerns about bias due to selectively available data as well as logistical challenges, such as the time required to redact and review extremely long records. In terms of whether to also seek confounder information during follow-up, initial analyses showed that the amount of follow-up time on drug before quitting or switching differs for saxagliptin and some of its comparators. Thus it might be valuable to seek information about reasons for cessation and in particular, to examine whether the reasons for cessation differ between saxagliptin and its comparators.

11. What steps can be taken to minimize missing data?

Sometimes, confounder information may be missing from phase 2 data sources (e.g., medical records). As discussed above in Section IV.C.3., some patients may not make regular visits for care, and not all providers order recommended labs or document relevant information. Several aspects of study design can help minimize missing data.

The overarching study design influences the likelihood of missing data. As discussed above, certain populations (e.g., people with diabetes) are more likely to have regular clinic visits, and quality measures promote standardization of lab testing and documentation. The use of a "new user" design with active comparators may help reduce problems with missing data, or at least differential missingness between the exposed and "unexposed" arms, because all study subjects are patients who initiate a new medication. The study team should consider how their study design is likely to affect the risk of missing data, given their specific context.

When selecting subjects for phase 2 data collection, additional inclusion criteria can be imposed to minimize missing data. For instance, the team could elect to review medical records only for patients with a certain level of utilization, e.g., two or more clinic visits in the 12 month period of interest. This approach has been taken in published studies, for instance requiring a total of four or more visits (over any time period).[59] One theoretical concern is how such a requirement would affect the sample and whether it might reduce generalizability. People with frequent healthcare visits may differ in important ways, including underlying health status, from people with no visits. People with frequent visits may be sicker, or alternatively, they may be health-seeking and as a result more adherent to medical recommendations. If there is concern that important information is being missed by excluding people with few or no clinic visits, though, the best solution is not to attempt to review their medical records, since little information will be gained. Instead, a different approach would be necessary, such as collecting information via questionnaire or interview.

Application to the saxagliptin example:

As discussed previously, several features of this study design will tend to reduce missing data: patients with diabetes receive frequent health care, and all patients in this study are new users of some medication. It would be reasonable to impose a utilization requirement, say two or more visits in the past 12 months in the setting or clinic from which the chart is being obtained (that is, the health care setting in which a provider initiated saxagliptin). The number of visits could be determined after preliminary review of the distribution of utilization for the patients that the team is considering sampling.

12. How lengthy a medical record review is reasonable or necessary?

The length of the medical record review has important implications for cost and efficiency. There is likely to be an inverse relationship between the length of time required for each chart review and the feasible phase 2 sample size. Reducing the sample size to allow for lengthy chart reviews could diminish the usefulness of the two-phase study, because it could lead to lower precision and greater uncertainty around final risk estimates. In the extreme, very small sample sizes in phase 2 could even introduce bias, as we demonstrate through simulations in Section V of this report. Thus, the investigative team should develop a chart review plan aiming to make reviews targeted and as brief as possible while still achieving the study's objectives.

Several factors will affect how long it takes to conduct medical record review. These include 1) the length of the medical record itself (which is strongly affected by the time period chosen as the focus for review); 2) the number of variables being collected; and 3) the complexity of the variables. Considerations relevant to selecting a time period were discussed earlier under Question 10. In terms of variable selection, we recommend that the investigative team focus on variables they have reason to believe could be strong confounders and resist the ever-present temptation to broaden the focus. Various statistical approaches can be used to estimate the strength of a potential confounder, which depends on its prevalence and the magnitude of its associations with outcome and exposure. Information may be available from the literature about the expected prevalence of the confounder and the strength of its association with the outcome, but usually we will not have information about its association with the exposure. While it is generally true that confounders with low prevalence will have less impact than those with higher prevalence, all other things being equal, it is certainly possible for relatively rare confounders to have a substantial impact. This can occur when they have particularly strong associations with the outcome and the exposure. Quantitative bias analysis can help shed light on how much bias might be produced by specific confounders under different assumptions, e.g., assuming a range of associations between that confounder and exposure and outcome status.

The potential benefit of collecting data about rare confounders in phase 2 also hinges on decisions that we have previously reviewed about study design and sampling strategy. Our discussion for Question 4 (above) concerned whether a two-phase study should stratify on and oversample based on a given (rare) confounder. If the decision has been made not to sample in this way, then the team should think through whether it is still worthwhile to seek information about that confounder in phase 2. If phase 2 sample size is small to moderate (e.g., 250 charts), and the confounder is expected to have prevalence of about 10%, then on average, only 25 people with that confounder will be sampled for phase 2, and these 25 will be spread across many strata (exposure x outcome x DP, at a minimum). We are not advocating dropping all uncommon confounders from the medical record review at this stage, but since the team may need to prioritize confounders, it is relevant to consider confounder prevalence as one factor.

Decisions about how much detail to collect about a confounder also have a major impact on the length of the review. These decisions can also be guided by existing literature, such as by what is known about the relationship of that confounder with the outcome of interest. For example, current smoking is a stronger predictor of cardiovascular risk than duration or intensity of prior smoking. Thus supplemental data collection could focus on identifying current smoking status at the time of medication initiation, rather than collecting detailed information about cumulative pack-years of tobacco exposure. In contrast, for hypertension, cardiovascular risk depends not only on the presence or absence of hypertension but on the degree of blood pressure elevation. Thus, rather than collecting data from the medical record about hypertension as a dichotomous variable, it would be preferable to record actual blood pressure values. Moreover, since a person's blood pressure can vary greatly over time (even within a single day), the collection of multiple values would be preferable to allow for a more accurate assessment of cardiovascular risk.

In summary, the desired length of the chart review will reflect the unique needs and context of the particular surveillance activity. A key consideration is that longer chart reviews generally translate into smaller phase 2 sample sizes, and these smaller sample sizes limit the ability of the two-phase study to provide meaningful information about the presence of a signal. Thus, keeping the chart review as

concise and focused as is reasonable will benefit the project. Piloting the chart review form will be an important early step, because it will demonstrate whether the actual review time is consistent with expectations and thus whether the sample size and data collection protocol are likely to be feasible given the available resources and project timeline.

Application to the saxagliptin example:

The list of potential confounders includes 10 health conditions or characteristics. Many have several facets and would require multiple questions on chart review. For example, there is interest in ascertaining presence/absence of hypertension as well as its duration and multiple blood pressure values. Several entries in Table 9 have multiple components or require the capture of multiple events or health conditions. For instance, to capture history of heart disease would require ascertaining 7 different events or conditions such as MI, CABG, angioplasty, and others. Many questions are relatively simple and ask for information that should be easy to locate (e.g., height, weight, lab values). If the chart review form were piloted and found to be too lengthy, then some aspects of the more complex variables could be omitted (e.g., number of diseased vessels found on cardiac catheterization; ejection fraction.) This list focuses primarily on variables that members of the two-phase study workgroup have personal experience ascertaining through medical record review and have found to be useful and feasible to collect.

D. ADDITIONAL EXAMPLE: INTRAVENOUS IMMUNOGLOBULIN (IVIG) AND RISK OF THROMBOEMBOLIC EVENTS

The two-phase studies workgroup discussed an additional example: surveillance examining risk of thromboembolic events (arterial events such as MI and stroke as well as venous events such as deep vein thrombosis [DVT]) in relation to use of intravenous immunoglobulin (IVIG). The reason for discussing a second example was to see if new considerations arose in addition to those identified through our discussions of the saxagliptin example. Thus we deliberately chose a second example that differed in many ways from the saxagliptin example: a different type of product (i.e. IVIG is a biologic agent, often given in the inpatient setting and sometimes used on an acute rather than long-term basis), with different indications. It is also used in a different patient population (people receiving IVIG could be but are not necessarily diabetic; many are acutely ill and hospitalized).

IVIG is a biologic product that consists of purified human immunoglobins (antibodies). It is used to treat a wide array of conditions, including immunodeficiency and inflammatory conditions such as Guillain-Barre syndrome, idiopathic thrombocytopenia purpura (an autoimmune disorder in which platelets are destroyed), and others. People receiving it for immunodeficiency may receive it periodically on a long-term basis, while those with acute conditions may receive it for a particular episode of illness or a disease flare. IVIG can also be used for prophylaxis after certain exposures, e.g., when an unvaccinated person is exposed to measles. Case series and small studies have reported that receipt of IVIG may be associated with thromboembolic events including MI, stroke, pulmonary embolism (PE) and DVT. The goal of this surveillance activity is to use Mini-Sentinel data to answer questions such as: what is the absolute risk of a thromboembolic event after receiving IVIG? What is the time course? Does the risk differ according to the specific product received, the dose, or the infusion rate?

The study population will include all people exposed to IVIG, to allow calculation of the absolute risks of the outcomes of interest, including in different time windows relative to exposure. Inference about the relative risk will come from studying only the people exposed to IVIG who experienced an outcome of interest within 4 weeks—that is, from a self-controlled case series. Risk of a thromboembolic event will

be assessed in an "exposed" time window (0-2 days after receipt of IVIG for arterial events, 0-13 days for venous events) and compared to a "control" time window when any acute increase in risk caused by IVIG exposure is expected to have waned. Primary analyses will use a "control" time window from 14-27 days after receipt of IVIG, and secondary analyses will explore the use of other "control" time windows. The reason for a self-controlled design is that it is expected to be very difficult to find an appropriate unexposed control group, particularly in terms of controlling for indication. That is, if a person with severe disease (e.g., Guillain-Barre Syndrome) is treated with IVIG, it would be difficult or impossible to find a comparable person who was not treated. The lack of a comparable external "unexposed" control group could lead to residual confounding. With the use of a self-controlled design, any characteristics that are fixed or change only over relatively long time periods will be the same during the "exposed" and "unexposed" time window. Thus much of the potential for confounding will theoretically be reduced.

There are challenges posed by the IVIG study that motivate supplementary data collection, specifically medical record review. As discussed above under Question 6, for biologic products, administrative data may not contain adequate detail about the exposure. Information such as the specific product, concentration, dose administered and infusion rate are not captured by administrative codes. Also, as discussed under Question 7, there is uncertainty about the relative timing of exposure and outcome. Some people will receive IVIG during a prolonged hospitalization, and the administrative data do not indicate on which day of the hospitalization the IVIG was received or which day the event occurred. In other cases, the event may occur on the same day as the exposure, and it will be necessary to review records to understand which occurred first. Also, the IVIG may be dispensed some time prior to actual administration, or enough IVIG for several courses of treatment might be dispensed on a single day for future use. The planned analysis requires information about whether the thromboembolic event occurred during the primary risk window, but it is difficult to determine this without more information about the timing of IVIG administration. Medical records may contain more detail about on which days the IVIG was actually administered and the timing of the outcome in relation to the exposure.

For a study with a cohort design, potential confounders of the association between IVIG and thromboembolic events would include: indication for IVIG use (and the severity of the indication); body mass index; smoking status; personal history of cardiovascular disease (CVD) and/or venous thromboembolism (VTE); family history of CVD or VTE; and hormone use including oral contraceptives and menopausal hormone replacement therapy. The choice of a self-controlled design should remove confounding by the majority of these characteristics (e.g., smoking status, personal and family history of CVD or VTE) because they are not likely to change over the 2 weeks separating the "exposed" and "control" time windows. Thus supplemental data collection is not felt to be needed to reduce confounding by any of these characteristics. There is some interest in collecting supplemental information about these characteristics so they can be considered as effect modifiers, conveying greater susceptibility to adverse outcomes. There are some potential confounders that could change rapidly over time, such as severity of the underlying indication, but these are expected to be difficult or impossible to measure from medical records and so there is no plan to collect supplementary information specifically to reduce confounding due to such factors.

Currently, the IVIG workgroup plans to conduct several hundred chart reviews with a focus on 1) validating outcomes; 2) clarifying the timing of the outcome in relation to IVIG exposure; and 3) collecting additional data about the IVIG exposure, such as specific product, dose received, etc. In many cases they will need to request and review 2 separate medical records for each affected person

because the exposure will occur in one clinical setting while the outcome will be diagnosed and treated in another. The IVIG workgroup does not yet know if there will be few enough events in IVIG-exposed people to allow medical record review for every such case. If there are too many cases to review, they will need to decide which cases to sample. Factors they may consider include the distribution of cases in different risk windows. For instance, they may wish to sample in a way that improves precision of estimates of absolute risks for specific time windows of interest.

In conclusion, our discussion of the IVIG example highlighted the need to consider the use of two-phase studies for purposes other than improved confounder ascertainment. This led us to incorporate more information in this report about the use of two-phase study designs to collect richer exposure information and to validate outcomes. This information is located in Section III of the report in subsection III.A.1. (Study Settings) and subsection III.B.1.a. Here in Section IV of the report, relevant information can be found in subsections IV.C.5, 6 and 7. The IVIG surveillance activity also provides an example where (in theory) confounding may be handled in large part through the study design itself. We should note that this study design will only be feasible for certain kinds of exposures and outcomes. When an outcome can be caused by an exposure at any time following initiation, and the medication is typically used chronically (e.g., saxagliptin), it will be difficult or impossible to identify a relevant "control" time period, and so other approaches (which may include supplemental data collection) will be needed to fully account for confounding.

E. RELEVANCE OF INITIAL SURVEILLANCE STUDY DESIGN TO THE DECISION TO PROCEED WITH A TWO-PHASE STUDY

Mini-Sentinel surveillance activities use a variety of study designs. The need for supplemental data collection may vary depending on the initial study design. As discussed above, the saxagliptin activity utilizes a cohort design in which new users of saxagliptin are compared to new users of active comparator drugs. In contrast, the IVIG activity uses a self-controlled case series design. The three modules for semi-automated surveillance in the Prospective Routine Observation Monitoring Program Tools (PROMPT) program include a self-controlled design and two cohort designs, one propensity-score matched and the other using more traditional adjustment methods.

As previously discussed, reasons for collecting supplemental data collection via a two-phase design include 1) to reduce confounding; 2) to gather more detailed exposure information; and 3) to collect additional outcome information, including to validate outcomes. Both of the cohort designs included in the PROMPT activity compare people exposed to a new therapy to other people who were not exposed (who may or may not have received an active comparator drug.) As in the saxagliptin example, this design raises the possibility of confounding, because there may be important differences between the exposed and unexposed groups. Thus for these designs, a two-phase study may be valuable for supplemental data collection to reduce confounding.

As discussed in Section IV.D., some surveillance activities lend themselves to use of a self-controlled study design, which may greatly reduce the need for supplemental data collection about confounders. Still, even in studies utilizing this design, supplemental data collection about outcomes or exposure may be desired. The concepts and strategies used in designing a two-phase study are relevant to this scenario as well, particularly when there are too many events to review all of them and thus the workgroup must select a sample for chart review. There is a strong motivation to focus medical record review on the most informative subjects, because of the cost of performing detailed medical record

reviews in terms of both money and time. The methods we have described for sample selection and for data analysis are relevant in this scenario, just as they are when the focus is on improving confounder adjustment.

F. RELEVANCE OF TWO-PHASE STUDY DESIGN TO THE USE OF SUPPLEMENTAL DATA AVAILABLE AT ONLY SOME DATA PARTNERS ("OPPORTUNISTIC" DATA)

Thus far, this report has exclusively focused on study designs in which subjects are deliberately sampled for phase 2 data collection based on characteristics available from phase 1 data. The assumption has been that phase 2 data require considerable resources to collect, and so it is important to improve efficiency by targeting the most informative subjects. In some cases, there may be supplemental data readily available from some DPs beyond what is in the Common Data Model – for instance, electronic data on laboratory results and vital signs. This is particularly true for integrated health care systems, where the same organization provides health insurance and care. We refer to these kinds of data as "opportunistic supplemental data" because for some DPs, the data are readily available and require relatively few resources to access. However, these data are not uniformly available for all subjects at all DPs. For efficiency, it may be desirable to utilize these existing data as fully as possible before considering new data collection. This scenario has features in common with the two-phase study design, but there are important differences. It is beyond the scope of this report to address all of the methodological challenges related to using opportunistic data. Here, we will point out similarities and differences between the two scenarios—a two-phase study with deliberate sampling and the use of opportunistic data—and describe additional methodological work that is needed to support the use of opportunistic data.

The most striking similarity between the two-phase study design and the opportunistic data approach is the existence of two stages of data collection. Both approaches first draw on automated data in the Mini-Sentinel Common Data Model for a larger population. They then obtain richer data for a smaller subset. Some of the statistical methods we discussed in Section III, such as reweighting, can be used to analyze data in both scenarios. However, these analytic techniques do not address the question of selection bias in the opportunistic sampling scenario, and so additional work is needed to overcome this challenge. In a two-phase study, selection is driven by the investigators, and so selection probabilities are known. There is far less concern about selection bias because people are deliberately sampled for new data collection. The team understands how people in whom the phase 2 data are available differ from those in whom they are missing, because sampling is deliberate and is based on characteristics measured in phase 1. Selection bias may still arise, for instance because of missing data (e.g., if medical records cannot be obtained for some study subjects.) Still, for the most part, it is possible to describe why the phase 2 data are available for certain people and not others. In contrast, in opportunistic sampling, it may not be clear why data are available for certain people and not others. One factor may be the health plan or health care system of which they are members, but there may be unmeasurable differences as well. Thus there are concerns about bias due to selective inclusion of only certain people in these analyses. Specific methodological strategies are needed to handle this bias. Investigating and discussing them is beyond the scope of this report. This topic will be addressed by a workgroup that will be constituted in late 2013 with the aim of exploring the use of laboratory and other data which are available for some DPs and not others.

Another difference between the deliberate and opportunistic sampling approaches is the need for a study design strategy. In a two-phase study, decisions must be made about which and how many people

to sample. A large portion of this report has focused on providing guidance for these decisions. In contrast, in opportunistic sampling, guidance about who to sample is not needed. Presumably, all available data will be used. Thus, material in this report about sampling strategies is not relevant to the use of opportunistic data. A key strength of the two-phase study design is the ability to target the most informative people, particularly those who are exposed to the medical product of interest and experience the outcome. The sample of people with opportunistic data may not happen to contain a large number or proportion of the most informative people. This disadvantage must be weighed against the advantages of obtaining data for a relatively large number of people with minimal extra effort.

To sum up, the two-phase study design and approaches using opportunistic data (available from some DPs but not others) have common features and also important differences. The study design considerations outlined in this report are not relevant to the use of opportunistic data, but analytic strategies outlined in Section III of this report may be relevant. Additional methodological work is needed to address issues related to the use of opportunistic data, including selection bias. There are plans within the Mini-Sentinel project for methodological issues related to the use of opportunistic data to be addressed by a separate workgroup in the near future.

G. A PROSPECTIVE APPROACH TO SUPPLEMENTAL DATA COLLECTION: BENEFITS AND DISADVANTAGES

Thus far, this report has focused on supplemental data collection after a signal has arisen. At that time, the population can be classified according to their exposure and outcome status, and a sample can be drawn targeting the most informative people (generally, those with both the outcome and exposure). There is an alternative approach that warrants consideration, which is to collect supplemental information for a sample of the population in a prospective manner, while initial surveillance analyses are still underway and before a signal has arisen.

This prospective approach to supplemental data collection has some theoretical advantages, including that it could dramatically shorten the time required to determine whether a signal is likely to be due to confounding, because detailed confounder information could have already been collected on a sample of the population by the time a signal arises. Mini-Sentinel projects which have reviewed medical records have often taken from 9 to 12 months to complete. This means that if supplemental data collection is needed, it could take 12 months or more to confirm or refute a signal. If the outcome is very severe, it may be felt that 9 to 12 months is too long to wait for more information. The desired time frame for confirming a signal may also depend on the exposure under study. In the case of influenza vaccine, a given vaccine is typically used only for a single influenza season. Thus, information about vaccine safety is potentially only of value for a relatively short time. If no definitive information can be obtained until the following flu season about 12 months later, a new vaccine will already be in use, and the value of the information obtained from supplemental data collection is greatly diminished. If supplemental data can be collected continually during initial surveillance, a definitive answer may be available in a more timely fashion.

Another theoretical advantage of prospective supplemental data collection is that in some cases, confounding could obscure a signal, meaning that surveillance might fail to detect a safety problem when one truly exists. In these scenarios, collecting better information early on through chart reviews could allow detection of a clinically important signal that would otherwise have been missed.

A prospective approach to supplemental data collection has been taken in the context of safety surveillance for vaccines, more specifically, within the CDC-funded Vaccine Safety Datalink. In the cases of which we are aware, supplemental data collection was desired to address concerns about poor measurement of outcomes of interest from administrative data. VSD investigators discovered that signals emerging during routine surveillance were sometimes found on further investigation not to be true signals because many of the apparent outcomes identified in administrative data were not valid. Thus, some VSD activities examining rare or high priority events carried out supplemental data collection prospectively to validate outcomes as they arose. This approach was used in a study of Guillain-Barre syndrome in relation to receipt of flu vaccine (personal communication, Eric Weintraub, Centers for Disease Control). Medical record review provided more information about the timing of the event in relation to receipt of vaccine and also shed light on site-specific coding practices that were causing signals to emerge at some sites but not others. Because Guillain-Barre syndrome is very rare, a two-phase study design was not needed; the group was able to review records for 100% of outcomes. Similarly, prospective medical record review was used to validate outcomes during active surveillance of Menactra examining risk of Guillain-Barre syndrome.

There are comparable situations within Mini-Sentinel where it might be desirable to collect supplemental data prospectively to validate outcomes, as discussed in Section B under Question 5. For instance, rhabdomyolysis is poorly measured by administrative data, with a positive predictive value reported to be as low as 8%. If misclassification is nondifferential, then it could lead to bias toward the null, obscuring a true safety signal. Prospective validation of this outcome could allow detection of a signal that might otherwise have been missed. The same considerations apply to other outcomes that may be poorly measured, such as severe acute liver injury.

A prospective approach to supplemental data collection has considerable disadvantages, however, which must be taken into account. Substantial resources are required, in terms of both money and time. The Mini-Sentinel program intends to eventually carry out routine surveillance for multiple exposure-outcome pairs at the same time. Some surveillance activities themselves include multiple comparisons of interest, e.g., in the saxagliptin example, new users of saxagliptin are compared to 4 different new-user comparator groups. Given the large number of exposure-outcome pairs that are of interest to Mini-Sentinel, it will not be practical to conduct prospective supplemental data collection for all pairs under surveillance at any given time. A relatively small group would need to be chosen. Waiting until a signal has emerged ensures that the resources required will be devoted to the right outcome-exposure pair.

Another consideration is that carrying out medical record reviews poses logistical challenges, particularly for the DPs who are not integrated health care delivery systems. Charts must be requested from outside health care providers, in some cases from multiple health care settings. Logistically, it may be more efficient to wait until a fairly large number of cases have been identified needing chart review, because it may be possible to request multiple charts from a given provider or health care system at the same time. This allows for more efficient training and scheduling of study team members such as people carrying out chart abstraction and redaction. Additional practical considerations that arise when considering a prospective approach to supplemental data collection include difficulty budgeting for chart review and contracting with DPs when it is not yet known how many records will need to be reviewed at each site.

In summary, there are some advantages and considerable disadvantages to collecting supplemental data on a prospective basis before a signal has arisen. While several activities have taken this approach, as described above, their primary focus was on validating rare outcomes that are poorly measured, and the relatively small number of outcomes meant that chart review posed a relatively lower burden in terms of time and money. Also of note, all of these activities involved Data Partners that are integrated health care delivery systems, at which medical record review is less challenging. The workgroup recognizes that within Mini-Sentinel, there may be some rare scenarios where a prospective approach is warranted. We expect that these will often involve rare outcomes that are poorly measured in administrative data that are of particularly high priority and are difficult to study via any other means. However, in general, it is likely that supplemental data collection via a two-phase study design will in most cases be considered only after a signal has arisen and extensive efforts have already been undertaken to investigate and understand the signal using other available resources.

H. SUMMARY

In this section, we have reviewed practical and logistical questions relevant to designing a two-phase study for supplemental data collection within Mini-Sentinel. We provided a list of questions that a workgroup should ask and answer (shown in Table 8), and we demonstrated the process by working through these questions for a specific example, the case of saxagliptin and MI.

For some questions that arise in designing a two-phase study, applying existing knowledge and logical reasoning may not be sufficient to provide a clear answer. The use of simulation studies can be a valuable next step. Thus in the next section of this report, we will describe and demonstrate the use of simulation studies to help inform two-phase study design in the context of Mini-Sentinel surveillance activities.

Table 7. Examples considered by the workgroup

Example	Population of interest	Comments
Saxagliptin and risk of myocardial infarction	Diabetics	Disease risk factors such as smoking, obesity, and diabetes severity may be confounders and may be poorly measured in electronic data.
Intravenous immunoglobulin and risk of thromboembolic events (arterial events such as myocardial infarction and stroke; venous events including deep vein thrombosis)	General	Some aspects of exposure (specific product, dose) are not available in electronic data but may be available in medical records. It would be useful to confirm the indication for use and assess the temporal sequence of exposure and outcome.
Dabigatran and risk of myocardial infarction or bleeding	Atrial fibrillation	Factors influencing prescribing, such as frailty and renal function, are hard to measure from administrative data.
Insulin and other diabetes medications and risk of cancer	Diabetics	Severity of diabetes and some cancer risk factors such as smoking and obesity might be confounders; these are hard to measure from electronic data. Note that cancer is a challenging outcome to study in the surveillance context given the long incubation period, meaning that people would need to be followed for a very long time after their initial medication exposure.
Tumor necrosis factor inhibitors and disease modifying antirheumatic drugs (DMARDs) and risk of cancer or infection	Rheumatoid arthritis	Severity of rheumatoid arthritis and/or inflammation is not well measured in electronic data.
Long-acting beta-agonists (LABAs) and risk of sudden death or exacerbation requiring hospitalization	Chronic lung disease	Severity of chronic lung disease may be a confounder and is difficult to measure from electronic data.
Various drugs and the outcomes of acute liver failure or acute renal failure	Varies	Outcomes themselves likely need validation from charts. Baseline kidney function is an important confounder, and many Mini-Sentinel Data Partners do not have access to electronic laboratory data.
HMG co-A reductase inhibitors (statins) and risk of rhabdomyolysis (comparing different statins or different doses)	General	Chart review is needed to validate the outcome and can also determine when other causes of rhabdomyolysis are present (such as trauma) so cases with these causes can be excluded.
Various drugs in relation to myocardial infarction and other cardiovascular outcomes	General	Disease risk factors such as smoking, obesity, and history of cardiovascular disease may be confounders and may be poorly measured in administrative claims data.

Table 7, continued. Examples considered by the workgroup

Example	Population of interest	Comments
Drospirenone-containing oral contraceptives and risk of venous thromboembolism	Reproductive-age women	Disease risk factors such as smoking, obesity, and family history of venous thromboembolism may be confounders and may be poorly measured in electronic data.

Table 8. Questions that must be answered in designing a two-phase study for supplemental data collection to improve measurement of and control for confounding

Identifying confounders of interest
1. What are potential key confounders?
Consider: indication for use, including severity of the condition that is an indication for use; disease risk factors; measures of frail health
What factors are likely to influence "early adoption" of this new drug or product?
2. Are potential confounders available in the phase 1 (electronic) data?
3. Can confounders be accurately measured from other sources (e.g., medical records)?
4. Should the phase 2 sampling scheme stratify on and oversample any confounders?
Other data that may be of interest
5. Does the outcome require validation?
6. Are there aspects of exposure that need to be measured from medical records?
7. Do records need to be reviewed to clarify the timing of the outcome in relation to the exposure?
Practical and logistical considerations
8. At which Data Partners should phase 2 data collection be carried out?
9. If medical records are to be reviewed, from which care setting and/or provider should these be obtained?
10. What time period should phase 2 data collection target?
11. What steps can be taken to minimize missing data?
12. How lengthy a medical record review is reasonable or necessary?

Table 9. Potential confounders of the association between saxagliptin use and acute myocardial infarction

Domain or category	Variable(s)
Demographic factors	Age, sex, race/ethnicity, socioeconomic status
Risk factors for the outcome, myocardial infarction	Hypertension (presence/absence, duration, blood pressure values)
	Body mass index (obtain height and weight)
	Smoking status (current/past/never)
	Hyperlipidemia (presence, cholesterol values)
	History of cardiovascular disease (including myocardial infarction, interventions such as coronary artery bypass grafting or angioplasty, current anginal symptoms, number of diseased vessels based on catheterization or imaging, history of stroke or peripheral arterial disease)
Measures related to indication for saxagliptin use, including severity	Diabetes severity: duration, hemoglobin A1c levels, insulin dose, presence of diabetic complications
Measures of frail health	Nursing home residence
Other measures	Congestive heart failure (presence; ejection fraction)
	Renal function (creatinine and/or estimated glomerular filtration rate)
	Cancer (including site and stage)

Table 10. Accuracy of administrative data for potential confounders for the saxagliptin surveillance activity*

Domain and variables	Comments on accuracy
Risk factors for MI	
Hypertension	ICD9 codes have moderate accuracy for identifying the presence of hypertension (e.g., sensitivity of 61%[54] in one study and 82% in another[55]) Medication data likely have good accuracy for identifying treated hypertension, although some medications may be used for other conditions which may lead to misclassification. Administrative data contain little useful information about duration or severity of hypertension. Severity can be measured using blood pressure values, which are available electronically from some DPs (integrated health care delivery systems with electronic medical records [EMRs]). However, blood pressure values are not available from the DPs contributing the most person-years of exposure, because for the most part these DPs are providing claims data.
Body mass index	ICD9 codes for obesity have very low sensitivity, e.g., 15% in a study of children.[56] Height and weight values are available from EMRs for some DPs but not those contributing the most person-years of exposure.
Smoking	ICD9 codes have low sensitivity (e.g., 7% in one study,[55] 32% in a second[60] and 38% in a third.[61])
Hyperlipidemia	ICD9 codes have moderate accuracy (e.g., sensitivity of 85% in one study[61] and 57% in another.[55]) Medication data likely have good accuracy for identifying treated hyperlipidemia, although these data may miss people with hyperlipidemia who are not being treated (e.g., due to medication allergies or intolerances.) Cholesterol values are available electronically from some DPs (those with EMRs) but not those contributing the most person-years of exposure (who typically supply claims data).
History of cardiovascular disease	Administrative data have high PPV for some procedures (e.g., CABG, PCI[53,62]) and diagnoses (e.g., PPV of 86% for acute MI in one study[10] and 78% in another with sensitivity 80%[62]), but sensitivity could still be limited because these events may have occurred prior to the baseline window during which confounders were defined, or even prior to health plan enrollment. Accuracy of administrative data is only fair for angina (e.g., in one study PPV was 40% and sensitivity 90%.[62]) Number of diseased vessels is not likely to be measurable from electronic data. For stroke, results of validation studies are extremely variable,[63] with studies reporting PPV for specific algorithms ranging from 31 to 90%[64] and sensitivity from 28[55] to 92%.[63] Electronic data have low accuracy for peripheral arterial disease (e.g., sensitivity 25% in one study.[65])

Table 10, continued. Accuracy of administrative data for potential confounders for the saxagliptin surveillance activity*

Domain and variables	Comments on accuracy
Measures related to indication for saxagliptin use, including severity	Administrative data have high accuracy for the presence of diabetes (e.g., sensitivity of 91% and specificity 99% in one study[55] and sensitivity 82%, specificity 97% and PPV 71% in another.)[65] They are less accurate for chronic complications of diabetes, e.g., sensitivity 49%[65] in one study and 59% in another.[66] Administrative data are likely to provide little information about diabetes duration especially if the baseline period during which covariates are measured is short. Hemoglobin A1c values are available electronically from some DPs but not those contributing the most person-years of exposure. Insulin dispensings are captured in pharmacy data, but dosing instructions are not included. It may be possible to estimate daily dose based on frequency of refills and amount dispensed.
Measures of frail health	Good to excellent accuracy for measuring receipt of nursing home care. In one study, sensitivity of claims data was 88% and PPV 84%,[67] while a second study reported sensitivity of 96.7% and specificity 99.9%.[68] Direct measures of functional status are not available. Claims for durable medical equipment such as home oxygen use or receipt of wheelchairs or hospital beds for home use may provide some information about functional status and debility. Comorbidity indices derived from administrative claims data may have limited ability to capture frail health,[45,48] as discussed in the text.
Other measures	
Congestive heart failure	Moderate accuracy for the presence of congestive heart failure, with one study reporting PPV of 45% and sensitivity 79%[62] and another reporting sensitivity of 77%.[61] Ejection fraction is not available through administrative data.
Renal failure/ chronic kidney disease	Measures of renal function (creatinine and/or estimated glomerular filtration rate) are available electronically from some DPs but not those contributing the most person-years of exposure. ICD-9 diagnosis codes are likely to have poor sensitivity for chronic kidney disease of moderate severity, e.g., chronic kidney disease stage 3 and 4, and higher sensitivity for very severe disease. Prior studies reveal that administrative data have low accuracy for renal disease, e.g., sensitivity 42% and PPV 63%.[65] Diagnosis and procedure data are likely to be highly accurate for identifying end stage renal disease requiring dialysis.
Cancer	Presence of cancer and cancer site are expected to be fairly well measured in administrative data. One study reported sensitivity of 70% and PPV 96% for presence of any malignancy and sensitivity and PPV both 88% for presence of a metastatic solid tumor.[65] Cancer stage may not be well measured in administrative data.

*This table is based on workgroup members' knowledge and past experiences, supplemented by information from other Mini-Sentinel investigators and reports as well as references from the literature. It is for illustrative purposes and is not intended to provide a systematic review of the accuracy of administrative data for these conditions. Its primary purpose is to demonstrate the thought processes needed to design a two-phase study for supplemental data collection about confounders.

V. THE USE OF SIMULATION STUDIES TO ANSWER DESIGN QUESTIONS RELATED TO TWO-PHASE STUDIES

A. INTRODUCTION

In previous sections, we reviewed methodological and practical considerations for conducting a two-phase study to investigate a signal arising from Mini-Sentinel surveillance activities. At several points we alluded to the possibility of using simulations to help guide study design choices for phase 2 supplemental data collection. In this section we provide more information about the usefulness of simulations for this purpose. We begin by briefly explaining what a simulation study is and how it is used in general to evaluate statistical procedures. Then we provide a detailed example demonstrating how in the context of Mini-Sentinel surveillance, one might use simulations to help answer questions such as:

1) How much information are we likely to gain from conducting a two-phase study? What is the expected amount of bias reduction, and how precise would estimates likely be?
2) How large a sample would be needed in phase 2 for it to be worthwhile to carry out such a study? Or in other words: if resources can support a phase 2 sample of a certain size, would it still be helpful to carry out supplemental data collection in this way?
3) We know that the sampling strategy should stratify on exposure and outcome; would it be valuable to also stratify and select patients based on additional characteristics, such as confounders?

As in Section IV, we use the saxagliptin surveillance activity to ground our discussion. We note though that, thus far, no signal has arisen and so this situation remains hypothetical.

B. SIMULATION STUDIES: GENERAL BACKGROUND

When planning a study that will employ some statistical procedure or estimator, it is important to understand how the procedure is likely to perform (e.g., in terms of bias, efficiency, power, etc.) in a range of hypothesized settings. Estimators arising from statistical methodologies that are commonly used in practice (e.g., binomial proportions, linear regression model coefficients and standard errors) often have formulas which can be used to compute operating characteristics of interest because these methodologies have been extensively investigated under a variety of settings. Other times, though, either because of the methodology or the complexity of the setting, formulas for the operating characteristics are not available. In these situations, simulations (also known as Monte-Carlo methods) provide an approach for examining the operating characteristics of the planned statistical procedure or estimator.

Such a simulation entails first running a computer program to generate a random sample of data conforming to the hypothesized setting under study. This requires that investigators make certain assumptions up front – for instance, in the Mini-Sentinel setting, a simulation would require assumptions about the sample size, the prevalence of the exposure and incidence of the outcome, the true relationship between exposure and outcome, and so forth. Next, the statistical methodology of interest is applied to the data, and estimated results are generated and stored. Then this process is repeated many times. At the end of the simulation runs (also known as trials), one can use the stored results (e.g., the estimates across the repeated random samples) to gain an understanding of how the statistical technique under review would be expected to perform in a specific setting.

In the Mini-Sentinel context, there is interest in understanding how accurately one can estimate the true association between the exposure and outcome and whether conducting a two-phase study to follow up on a signal can meaningfully improve the validity of the estimate. In a simulation, the true association is known because it has been specified as part of the assumptions made ahead of time. Thus, by comparing the estimated exposure/outcome association from the statistical model across the many simulation trials to the "true" association specified as a simulation input, one can determine both the accuracy and variability of the model under the hypothesized settings (e.g., different two-phase study designs). By varying the assumptions used in the simulation, one can investigate the performance of the two-phase approach in different settings (e.g., with small sample sizes, rare outcomes, etc.)

C. SIMULATION STUDIES: APPLICATION TO PLANNING FOR A TWO-PHASE STUDY WITHIN MINI-SENTINEL

As noted above, simulation studies generate random samples of data that are intended to conform to the setting being investigated. That is, the simulated data reflect some hypothesized reality. The simulation study thus requires certain inputs (parameters) that will define the true underlying relationships that govern the data-generating mechanism. This means that the surveillance team needs to specify, up front, what the true distribution and relationships within the data are likely to be. A range of values can be provided for these parameters so that the simulation results can provide insight into the performance of the statistical method across a variety of settings.

Table 11 lists the types of parameters that would likely need to be specified for most simulation studies to help guide decisions about conducting a two-phase study within Mini-Sentinel.

Table 11. Parameters needed to conduct a simulation to help plan a two-phase study

Sample size (size of the study sample) in phase 1
Exposure prevalence
Prevalence of potential confounder(s) of interest
Strength of confounding: an estimate of the confounder-exposure association(s), and an estimate of the confounder-outcome association(s)
Expected incidence of the outcome
Underlying (true) association between the exposure and outcome, e.g., null (RR=1.0) or direction and magnitude of the association
Sample size for phase 2 data collection
Two-phase estimation method to be employed (e.g., weighted, profile, or maximum likelihood)

1. Application to saxagliptin example

To demonstrate the utility of simulations in planning a two-phase study, we considered a hypothetical example. In this example we posited that Mini-Sentinel routine surveillance activities detected a potential signal indicating increased risk of myocardial infarction (MI) associated with saxagliptin use relative to use of the comparator drug, sitagliptin. Because this hypothetical signal is based on administrative data with limited ability to measure potential confounders, supplemental data collection (via medical record review) might be considered to determine whether the apparent increased risk of MI is likely to be due to saxagliptin use (a true causal association) or whether the association reflects unmeasured confounding.

A simulation prior to phase 2 data collection could help the team leading surveillance explore the potential usefulness of this approach to supplemental data collection under different plausible realities. The conditions we examined (i.e., the simulation parameters) are shown in Table 12.

When selecting parameters for potential confounders, we had smoking in mind when we considered confounder 1 and obesity for confounder 2. We recognize that in reality, there are likely to be more than 2 confounders of interest. It is not practical in a simulation study to include a large number of confounders, so we tried to address this constraint by increasing the strength of confounding. We accomplished this by specifying a relatively strong association between each confounder and the exposure and each confounder and the outcome. The resulting associations are perhaps stronger than would be expected in reality for individual confounders, but instead, one might think of these confounders as representing the net effect of many confounders.

Table 12. Inputs for saxagliptin simulation

Sample size	150,000 users of saxagliptin or sitagliptin
Exposure prevalence	20% saxagliptin, 80% sitagliptin
Prevalence of confounder(s)	10% confounder 1 40% confounder 2
Strength of confounding (confounder-exposure OR)	OR = 3.00 (confounder 1-saxagliptin) OR = 2.00 (confounder 2-saxagliptin)
Strength of confounding (confounder-outcome OR)	OR = 4.00 (confounder 1-MI) OR = 2.00 (confounder 2-MI)
Expected incidence of the outcome, MI	1/100
Underlying (true) association between exposure and outcome	OR = 1.00 (saxagliptin-MI)

We assumed that no measures were available at phase 1 for these two important confounders (i.e., smoking, obesity) and that no true relationship between saxagliptin and MI existed (OR=1.00). We also assumed that phase 2 sampling strata were to be formed on the basis of exposure and outcome and that exposure and outcome were both known with complete accuracy at phase 1.

We began by considering what would happen if we were to sample 1000 people for medical record review using a balanced design, i.e., equal numbers of individuals selected from each stratum. We first present this example in detail and then present results from additional simulations assuming smaller phase 2 sample sizes (500, 250 and 100 medical record reviews).

Table 13 (next page) shows the expected distribution of the phase 1 data according to these exposure and outcome parameters, as well as the corresponding sampling fractions for the phase 2 sample, for the scenario where 1000 medical records are to be reviewed. Note that at phase 1 the confounder status is not known, because we assume the confounders are not measurable from the administrative data (thus the need for phase 2 data collection).

Table 13. Sampling scheme for simulation with 1000 patients sampled for phase 2

MI	Saxagliptin Use	N	To be sampled at phase 2	Sampling fraction
No	No	118,894	250	0.002
No	Yes	29,606	250	0.008
Yes	No	1105	250	0.226
Yes	Yes	395	250	0.633

Assuming this expected distribution of phase 1 data, the initial routine surveillance analysis based on data in the Mini-Sentinel common data model (with no confounder information, as we have hypothesized) would yield an OR of 1.44 (95% CI: 1.28, 1.61), indicating an elevated risk of MI associated with saxagliptin use. Thus, for the simulation setting we have hypothesized, the phase 1 data would be expected to provide a "signal" of possible increased risk of MI with saxagliptin use.

2. Results for Simulation 1, assuming 1000 medical record reviews

We ran 10,000 trials of the simulation. Each time we:

1) Generated a random phase 1 sample of size 150,000 according to the set parameters governing the relationships between exposure (saxagliptin use), outcome (MI), and confounders (C1 and C2);
2) Stratified the phase 1 sample on the basis of saxagliptin use and MI;
3) Generated a random, balanced phase 2 sample of size 1000 from the above phase 1 strata, with the goal of sampling 250 patients per strata;
4) Recorded the true confounder measures (C1 and C2) for the 1000 sampled individuals;
5) Used a two-phase estimation method (weighted likelihood as described in Section III.C.1) to estimate parameters of a logistic regression model for the saxagliptin-MI relationship, adjusted for C1 and C2, using this sample; and
6) Stored the estimated adjusted odds ratio (OR) for the saxagliptin-MI relationship and the corresponding 95% confidence interval (CI). This allows us to summarize the expected performance of a two-phase study in the hypothesized setting.

A key point is that while we utilized a *weighted likelihood*-based two-phase estimation approach for this simulation example (and for all examples in this section), in practice researchers would benefit from comparing the performance of different estimation methods (e.g., weighted, profile, and maximum likelihood) for their hypothesized setting. There may be important tradeoffs between bias and precision across these methods, especially in settings with small samples and the possibility of model misspecification. As this section of the report is simply meant to illustrate how a simulation tool might be used in planning a study, we did not examine such comparisons.

Figure 2 (next page) shows a histogram of the distribution of OR estimates generated by these 10,000 simulations, representing the results of potential two-phase studies for the saxagliptin-MI association. The figure shows that, on average, the estimated ORs across simulated trials center around 1.00 – meaning that the two-phase study methodology is providing an unbiased estimate (since the average OR is the same as the truth, which was set as an input parameter for this simulation). The histogram also shows that variability in the OR across simulation trials is relatively small: the histogram has a high peak

and is relatively narrow. The variability is small in part due to the choice of a large phase 2 sample size (N=1000 patients selected for medical record review).

Figure 2

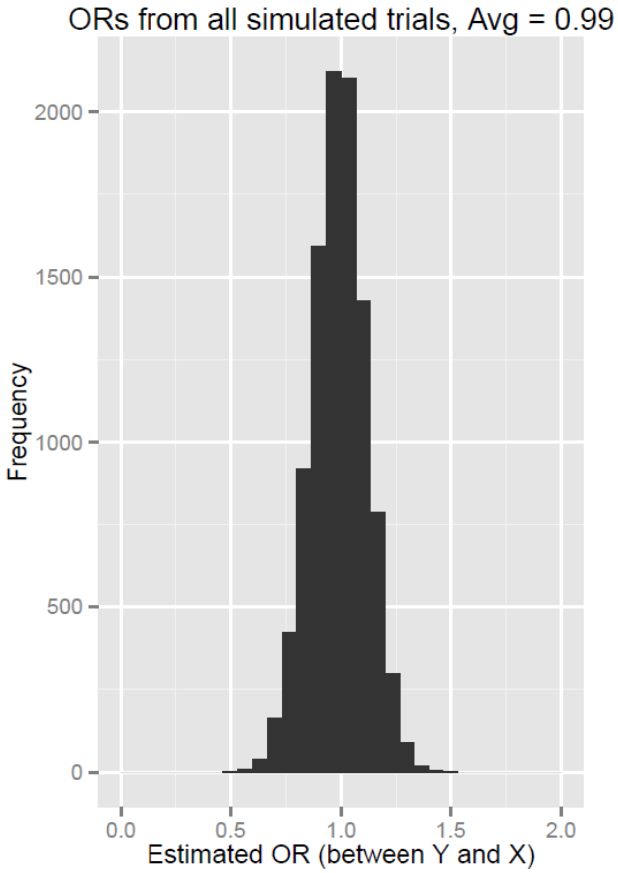


Figure 3

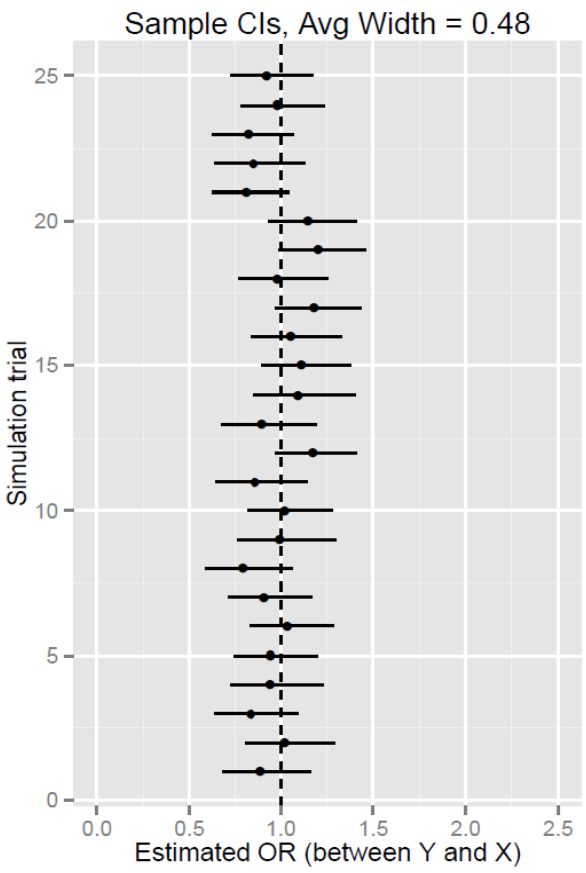


Figure 3 shows a sample of the ORs and 95% CIs generated by the simulation runs. For practical reasons we present results from 25 of the 10,000 simulated trials. This helps illustrate, in a different manner than Figure 2, the amount of variability in the estimates using this phase 2 sample size and design. Note that precision is quite good, with CIs having an average width of 0.48 (on the odds ratio scale, as shown in Figure 3).

Finally, Table 14 (next page) shows how often the 95% CI constructed for the estimated OR of interest excluded a particular value. Note that the true OR is posited to be 1.00 in this simulation, so if the 95% CIs have the proper coverage level (that is, if they are truly 95% confidence intervals), then they should only exclude 1.00 (the null) approximately 5% of the time (i.e., $\alpha = 0.05$). Table 14 shows that in this case, the 95% CI does provide the appropriate level of coverage.

Table 14 also shows that most (89%) of the two-phase study simulation trials would provide a 95% CI excluding an OR of 1.4, and a majority (64%) would also exclude a value of 1.3. Remember that for this example, the analysis using only phase 1 data would, on average, have yielded an OR of 1.44 for the MI-

saxagliptin association, with CIs having an average width of 0.33 (on the odds ratio scale). The results of this simulation suggest that if the true relationship is actually null (OR=1.0) and the spurious initial signal is being driven by unmeasured confounding (as specified in the simulation parameters), then a phase 2 study that samples 1000 charts has a strong likelihood of leading investigators to conclude correctly that there is no true association between saxagliptin use and MI risk – or at least that the association, if true, is considerably smaller than initially thought. Assuming the simulated framework accurately reflects reality, the OR estimate and corresponding CI that results from such a phase 2 study will likely contain the null, and the confidence interval would be fairly narrow.

Table 14. Proportion of CIs excluding a given OR in simulation trials with phase 2 sample size of 1000

Odds ratio	% of CIs excluding the specified OR
1.0	5
1.1	9
1.2	31
1.3	64
1.4	89
1.5	98

3. Results from simulations varying the size of the phase 2 sample

Since reviewing 1,000 medical records may not be routinely feasible in Mini-Sentinel, we examined the impact of sampling fewer individuals in phase 2, specifically, sampling 500, 250 or 100 patients for medical record review. Past Mini-Sentinel projects that have reviewed medical records to validate outcomes have carried out approximately 100 to 250 reviews over 9 to 12 months. Thus, a sample size of 250 is probably more realistic than 1,000 given the logistical challenges of reviewing medical records across multiple Data Partners and health care systems.

In this section, we present results from simulations under the same settings as above (described in Table 12, Section V.C.1.) but assume phase 2 sample sizes of 500, 250 or 100 medical record reviews. We compare these results to the results derived with a sample size of 1000 (shown in Section V.C.2.) We present results graphically and also describe and tabulate the findings.

Table 15 is an extended version of Table 13. It shows the number of patients who are expected to be sampled from each of the exposure-outcome strata using a balanced design under different phase 2 sample sizes.

Table 15. Sampling scheme for simulations with different phase 2 sample sizes

			To be sampled at phase 2			
MI	Saxagliptin Use	N	1000	500	250	100
No	No	118,894	250	125	62	25
No	Yes	29,606	250	125	62	25
Yes	No	1105	250	125	62	25
Yes	Yes	395	250	125	62	25

In Figure 4 at the end of this section, we present the histograms for the ORs resulting from each simulation, varying the phase 2 sample size. The first panel (far left) shows the results for a sample size of 1000 patients in phase 2, while the second panel (immediately to the right) shows the results when 500 patients are sampled in phase 2. With 500 patients sampled, the average OR is 0.98, still very close to 1.00 (the parameter we set for the simulation). Note that the histogram is somewhat wider now, reflecting greater variability in estimates when a smaller phase 2 sample is drawn.

The third and fourth panels show results when 250 and 100 patients, respectively, are sampled at phase 2. In addition to greater variability in OR estimates across trials, the histograms also reveal that samples yielding an estimated OR well below the truth are occurring with greater frequency than occurred with larger phase 2 samples. This is particularly noticeable for phase 2 samples of size 100. These extreme OR estimates are leading to a downward bias: the average ORs for phase 2 samples of 250 and 100 patients are 0.95 and 0.87, respectively. We explored the data further to understand why such extreme ORs were being generated. We found that the very low ORs resulted from phase 2 samples that included very few patients with some confounder/exposure/outcome combinations. These samples had such sparse information about the underlying confounder(s) that the resulting estimates were unreliably low. This is a 'sparse data problem' that we expect would be identified by a skilled analyst if it arose in practice, and the results would be readily identified as unreliable. Still, by that time substantial resources would already have been invested in collecting phase 2 data on that particular sample, and the ultimate result would be that the supplemental data collection would not have helped clarify the meaning of the original signal.

The reason for the high prevalence of these 'sparse' samples when phase 2 sample size is 100 can be illustrated by reconstructing the original Table 13 (p. 60) to show not only the expected distribution of the phase 1 data but also the expected confounder distributions (the confounders that are not measurable at phase 1) and the expected counts for various combinations of confounders, exposure, and outcome in the phase 2 sample. This reconstructed table is shown below as Table 16 (next page).

Given such small expected counts at phase 2 in many strata (last column), it is not surprising that some samples have extremely small or zero cells for certain exposure/outcome/confounder combinations, which in turn can make estimation from the logistic regression model unreliable.

Finally, we note that if we excluded some of the most extreme OR estimates (those based on these sparse samples, making up about 3% of the trials for the 250 patient sample and 29% for the 100 patient sample), the bias in the OR was reduced (average OR=0.96 for both settings). Still, this investigation showed that there is a fairly high chance that a phase 2 sample of 100 would not include enough confounder information to produce accurate estimates of the true exposure-outcome association– in essence, defeating the purpose of the two-phase study.

Figure 5 at the end of this section presents examples of the ORs and 95% CIs generated by individual simulation trials under differing sample sizes for phase 2. It tells a similar story as Figure 4. As the sample size in phase 2 decreases, there is more variability in the estimates derived from the simulation trials. Precision becomes worse (CIs are much wider). As discussed above, results with a phase 2 sample size of 100 (shown in the last panel of Figure 5) include quite a few trials with ORs different from 1.0 (though many still have CIs which include the null). As we have discussed, it would likely be difficult to draw firm conclusions from such estimates.

Table 16 (revised version of Table 13). Sampling scheme for simulation with 100 patients sampled for phase 2, showing the expected number of patients with each combination of exposure, outcome, and confounders in the final phase 2 sample when only exposure and outcome status are known at phase 1

MI	Saxagliptin Use	Confounder 1	Confounder 2	N	Known at phase1*	To be sampled at phase 2	Expected final phase 2 sample
No	No	No	No	69,490	118,894	25	15
No	No	Yes	No	5,958			1
No	No	No	Yes	40,508			9
No	No	Yes	Yes	2,938			1
No	Yes	No	No	11,060	29,606	25	9
No	Yes	Yes	No	2,845			2
No	Yes	No	Yes	12,895			11
No	Yes	Yes	Yes	2,806			2
Yes	No	No	No	388	1,105	25	9
Yes	No	Yes	No	133			3
Yes	No	No	Yes	453			10
Yes	No	Yes	Yes	131			3
Yes	Yes	No	No	62	395	25	4
Yes	Yes	Yes	No	64			4
Yes	Yes	No	Yes	144			9
Yes	Yes	Yes	Yes	125			8

*Only exposure (saxagliptin use) and outcome (MI) are known at phase 1, and thus formation of strata for selecting the phase 2 sample cannot utilize information on the confounders.

This simulation demonstrates that several issues can occur when a relatively small phase 2 sample is drawn: 1) bias may arise if there are too few phase 2 subjects with the relevant confounder(s) to allow adequate adjustment; 2) confidence intervals may be too wide to provide us with adequate certainty in our conclusions; 3) standard analyses, like the logistic regression analyses performed in this simulation, may fail due to sparse data. The sample size at which each of these issues causes a problem depends on the specific setting and is influenced by the true underlying relationships between exposures, outcomes, and confounders. Problems 1 and 2 occur because there is simply "not enough" phase 2 information and can only be addressed by increasing the phase 2 sample. With sparse but "enough" data, problem 3 could be addressed by using exact statistical methods. The extent to which the results of our current simulation reflect problems related to an overall lack of information versus problems specific to the use of standard statistical methods (logistic regression) rather than exact methods was not studied but warrants further investigation.

Table 17 (next page) provides another look at how bias and precision change as the phase 2 sample size decreases. It shows the proportion of CIs that would exclude the null as well as certain other values for the OR for the saxagliptin-MI association. We show these proportions for each phase 2 sample size that we examined in this simulation study: 1000, 500, 250 and 100 patients. As a reminder, the "true" OR

posited in this simulation was 1.00, and the spurious result generated from the phase 1 (administrative) data because of confounding was 1.44 (see Section V.C.1.)

Table 17. Proportion of confidence intervals (CIs) excluding a given odds ratio (OR) in simulation studies with different phase 2 sample sizes*

	% of CIs excluding the stated OR, for a phase 2 sample of size:			
Odds ratio	**1000**	**500**	**250**	**100**
1.0	5	6	8	15
1.1	9	6	7	13
1.2	31	14	10	13
1.3	64	35	16	14
1.4	89	63	26	17
1.5	98	85	45	21

*Note that all simulations posit a "true" OR of 1.0 for the exposure-outcome association; see Section V.C.1 for simulation parameters.

Table 17 shows the excellent performance of the two-phase design when either 1000 or 500 patients are sampled in phase 2. We have previously shown that in both cases, the two-phase study design does a good job of removing the confounding bias: average ORs are 0.99 and 0.98 respectively (Figure 4), contrasting with the average confounded estimate of 1.44 from an initial analysis using only phase 1 data. In addition, with phase 2 sample sizes of either 500 or 1000, the confidence intervals have (approximately) the proper coverage probability– that is, only about 1 in 20 studies would falsely exclude the null (consistent with a type 1 error rate or alpha set at 0.05; see Table 17, first row).

In contrast, as the sample size becomes smaller, the two-phase methods begin to break down. First, the 95% confidence intervals no longer provide appropriate coverage when the phase 2 sample size is 250 or 100. Instead of 5% of intervals excluding the null, approximately 8% and 15% of simulated trials, respectively, generated a CI that excluded the null. This improper coverage is likely a consequence of the small sample bias and the more frequent "sparse" samples noted above. Even after excluding the extreme OR estimates from these sparse samples in our simulations (as was done when examining bias above), the coverage probability of the null (OR 1.0) was still a little too low (7% and 8%, respectively).

Additional conclusions can be drawn from Table 17 about how the expected precision of the OR estimate changes in relation to phase 2 sample size. Recall that the true OR for the saxagliptin-MI association was posited to be 1.00 and the expected confounded estimate from only phase 1 data was 1.44. The goal in conducting a two-phase study is to remove bias, and so ideally we would like to derive estimates with 95% CIs narrow enough to rule out a moderately strong positive association between saxagliptin and MI in this setting. When a phase 2 sample of 1000 is chosen, 98% of simulation trials had 95% CIs narrow enough to exclude ORs of 1.5, 89% excluded an OR of 1.4, and 64% excluded an OR of 1.3. These results provide considerable evidence that the true OR is close to the null, and a reasonable interpretation would be that an initial OR of 1.44 was caused by residual confounding, rather than a true causal association. When the phase 2 sample size is 500, there is still a relatively high likelihood that the study would be able to rule out an OR of 1.4, though it becomes much less likely that an OR of 1.3 could be ruled out. The CI has become wider due to the lower sample size. When phase 2 sample size is

reduced to 250, it becomes even more difficult to rule out an elevated OR: now there is only a 45% chance that the phase 2 study would generate a CI excluding 1.5, and the majority of simulation trials with this phase 2 sample size would not be able to exclude an OR of 1.4 – the magnitude generated by the initial confounded analysis using only phase 1 data. Results with a phase 2 sample of only 100 patients show even worse precision (in addition to the concerns discussed above.)

Thus, through this simulation study, we were able to determine that in this particular setting, given parameters based on the saxagliptin example and assumptions about the true governing relationships between saxagliptin use, MI, and the confounders, a phase 2 sample size of either 500 or 1000 would be likely to provide useful results. It would be effective at reducing bias while maintaining good to excellent precision. It is very likely that a phase 2 study with this sample size would yield results that would change our interpretation of the initial findings in this hypothetical saxagliptin surveillance activity. A sample size of 250 would provide some useful information (the OR would be much closer to the null), but there would be some difficulty in interpreting results, and CIs would often be too wide to rule out an OR of the magnitude that provided the initial signal. In contrast, a phase 2 sample size of 100 would yield very imprecise estimates, and there is real potential that it would generate an OR estimate considerably different than the null. There is a genuine possibility, too, that a sample of this size would be too small to reasonably estimate a confounder-adjusted OR – thus defeating the purpose of the two-phase study.

4. Additional explorations using the saxagliptin example

In the prior section (Section V.C.3.), we examined the impact of varying the phase 2 sample size on the performance of a two-phase study estimating the saxagliptin-MI association. In this section, we evaluate performance when several additional parameters vary: the nature of the true association between saxagliptin and MI, the strength of confounding, and the prevalence of the outcome of interest. We also explore the impact of stratifying on a rare confounder when selecting the phase 2 sample.

a. Assuming a true positive relationship between the exposure and outcome

Our initial simulations posited a null association between saxagliptin and risk of MI (true OR=1.00). The next set of simulations assumed that a true association was present between saxagliptin and risk of MI, with an OR of 1.5. We kept most parameters the same as in the prior Table 12 (Section V.C.1.) but changed the nature of the confounding such that the apparent OR (from initial analyses using administrative data only) would be expected to be closer to the null – that is, confounding would bias findings towards the null. Table 18 (next page) shows the inputs for this simulation, with changes shown in bold.

With the settings shown in Table 18, a logistic regression analysis using only phase 1 data would, on average, yield an OR for MI associated with saxagliptin use of 1.34. The average CI width would be 0.31 on the OR scale, so for example an OR of 1.34 would have a 95% CI of 1.19-1.50. Thus again the initial surveillance activity would be expected to yield a signal, and it would be of similar magnitude as that seen in our first set of simulations.

We again examined the impact of varying the phase 2 sample size. Detailed results for the extreme cases (phase 2 sample sizes of 1000 and 100) are shown in Figure 6. Notable findings were that, in this

Table 18. Inputs for saxagliptin simulation, now with a true association present

Sample size	150,000 users of saxagliptin or sitagliptin
Exposure prevalence	20% saxagliptin, 80% sitagliptin
Prevalence of confounder(s)	10% confounder 1 40% confounder 2
Strength of confounding (confounder-exposure OR)	OR = 3.00 (confounder 1-saxagliptin) OR = 2.00 (confounder 2-saxagliptin)
Strength of confounding (confounder-outcome OR)	**OR = 0.40** (confounder 1-MI) **OR = 0.80** (confounder 2-MI)
Expected incidence of the outcome, MI	1/100
Underlying (true) association between exposure and outcome	**OR = 1.50** (saxagliptin-MI)

case positing a true exposure-outcome association, all sample sizes resulted in a negligible amount of bias. For example, even with a phase 2 sample size of only 100, the average OR from 10,000 simulation trials was 1.46 – very close to the assigned parameter value of 1.5. Furthermore, we did not encounter the problems with sparse phase 2 samples that arose in the last set of simulations. Overall, across the range of sample sizes for phase 2, there was always a high likelihood that CIs would exclude the null – that is, the results from the two-phase study would (correctly) confirm the finding of a higher MI risk with saxagliptin. For instance, with phase 2 sample size of 1000, 87% of simulation trials generated CIs excluding 1.2 or lower. Even with a phase 2 sample size of only 100, there was still a 64% chance the CI would exclude 1.0.

In summary, the two-phase design performed better when there was a stronger (positive) association between the exposure and the outcome, with little bias and greater precision.

b. Assuming stronger confounding is present

We repeated our simulations assuming stronger confounding was present. For this simulation we returned to the original simulation parameters, including assuming that there was no true association between the exposure, saxagliptin, and outcome, MI. Parameters are shown in Table 19 below, with changes from the initial simulations shown in bold.

Table 19. Inputs for saxagliptin simulation with stronger confounding

Sample size	150,000 users of saxagliptin or sitagliptin
Exposure prevalence	20% saxagliptin, 80% sitagliptin
Prevalence of confounder(s)	10% confounder 1 40% confounder 2
Strength of confounding (confounder-exposure OR)	OR = 3.00 (confounder 1-saxagliptin) OR = 2.00 (confounder 2-saxagliptin)
Strength of confounding (confounder-outcome OR)	**OR = 8.00** (confounder 1-MI; formerly 4.00) **OR = 4.00** (confounder 2-MI; formerly 2.00)
Expected incidence of the outcome, MI	1/100
Underlying (true) association between exposure and outcome	OR = 1.00 (saxagliptin-MI)

With these settings, a phase 1 data only analysis using only phase 1 data would be expected to yield a stronger signal than the initial simulation, on average an OR of 1.87 (95% CI 1.68-2.09) for MI risk in relation to saxagliptin use, even though the true OR (per the simulation settings) is 1.0. Because our initial simulation showed that a phase 2 sample size of 100 performed poorly, in these simulations we explored phase 2 sample sizes from 1000 to 250.

We found that with stronger confounding present, there was still little bias in the ORs for phase 2 sample sizes of 1000 and 500, but more bias was present for a sample of 250 than in the initial simulation setting. The average OR from simulation trials with a phase 2 sample of 250 was 0.88 in the setting with stronger confounding, compared to 0.95 in the initial setting. Furthermore, variability was notably increased for all phase 2 sample sizes, leading to less precise confidence intervals. For example, with a phase 2 sample size of 500, the average width of the CIs was 1.01 on the OR scale (compared to 0.64 in the initial simulation study). The coverage of the 95% CIs was less accurate than before, too: with a sample size of 500, 9% of CIs would exclude the null, compared to only 6% when confounding was weaker. (Recall that with an alpha level set at 0.05, only 5% of CIs should exclude the null if coverage levels are accurate.) Furthermore, the problem of some simulated samples having too sparse information on confounders began to appear regularly with a phase 2 sample size of 250 (approximately 36% of the time), a problem which did not occur as extensively at this sample size in the original simulation scenario. After excluding the extreme OR estimates based on these sparse samples, bias was no longer present for the sample size of 250, but variability was still extremely high with an average CI width of 1.37 on the OR scale (compared to 0.87 in the original simulation setting.) It is notable that with very strong confounding present, a sample size of 250 performed at least as poorly as, and perhaps worse than, the smallest phase 2 sample size (N=100) in the initial simulation runs.

In summary, under these assumptions, the two-phase study design performed more poorly when stronger confounding was assumed to be present. To avoid bias and improve the precision of estimates, a larger sample size would be needed in phase 2.

c. Assuming the outcome is more rare

We repeated our simulations assuming a lower incidence rate for the outcome, 1 in 1000 instead of 1 in 100. For this simulation we retained all other initial simulation parameters, including assuming no true association between the exposure and outcome. Parameters are shown in Table 20 below; the changes from the first set of simulations are shown in bold.

Table 20. Inputs for saxagliptin simulation with a more rare outcome

Sample size	150,000 users of saxagliptin or sitagliptin
Exposure prevalence	20% saxagliptin, 80% sitagliptin
Prevalence of confounder(s)	10% confounder 1 40% confounder 2
Strength of confounding (confounder-exposure OR)	OR = 3.00 (confounder 1-saxagliptin) OR = 2.00 (confounder 2-saxagliptin)
Strength of confounding (confounder-outcome OR)	OR = 4.00 (confounder 1-MI) OR = 2.00 (confounder 2-MI)
Expected incidence of the outcome, MI	**1/1000** (formerly 1/100)
Underlying (true) association between exposure and outcome	OR = 1.00 (saxagliptin-MI)

With these settings, a phase 1 analysis would yield, on average, an OR of 1.43 for MI risk in relation to saxagliptin use, with an average CI width of 1.07 (on the OR scale): for example, for an OR of 1.43, the 95% CI would be (0.99, 2.06). Again we explored phase 2 sample sizes from 1000 to 250.

Table 21 shows the distribution of exposure and outcome status in the underlying cohort and how each stratum would be sampled under different phase 2 sample sizes. Note that for any of the phase 2 sample sizes, there are not enough cases to "fill up" the planned 250 from each cell, and so the sample contains as many cases as are available for a given cell and then "fills up" the remaining available slots with controls. All phase 2 sample sizes, even N=250, include 100% of exposed cases, and sample sizes of 500 or 1000 capture 100% of unexposed cases as well.

Table 21. Sampling scheme with different phase 2 sample sizes when the outcome is more rare

			To be sampled at phase 2		
MI	**Saxagliptin Use**	**N**	**1000**	**500**	**250**
No	No	119,890	426	176	70
No	Yes	29,962	426	176	70
Yes	No	109	109	109	70
Yes	Yes	39	39	39	39

Compared to the initial simulation, performance of the two-phase design with regard to bias changed little when we assumed a more rare outcome. For example, for a phase 2 sample size of 500, the average OR from all simulated trials was 0.99, practically the same as in the initial simulation, and for a sample size of 250, the average OR was 0.97. (Recall that the "true" OR posited in the simulation is 1.0.) Precision, however, was affected. Confidence intervals were wider; average CI width was 0.64 (on the OR scale) in the original simulation vs. 0.94 with the rarer outcome for a phase 2 sample size of 500. The coverage of the 95% CIs was fairly similar regardless of outcome prevalence.

In summary, even with a rarer outcome, the two-phase study design still provided excellent bias reduction, but precision was substantially lower. The wider CIs that result could in some cases make it difficult to draw firm conclusions about the meaning of the initial signal, and as a result a larger phase 2 sample might be preferred. Simulation studies such as those presented here can help shed light on the expected precision with various phase 2 sample sizes in a specific setting when a two-phase study is being considered.

d. Stratifying on a rare confounder

In Section IV.C.4. (pp. 33-34), we discussed the potential benefits of stratifying on a phase 1 confounder (or its proxy) and oversampling on it for phase 2 data collection. In this simulation, we explored the impact of such stratification on the performance of the two-phase design. In particular, we were interested to see if this oversampling could solve the problems we observed in the simulations described above (Section V.C.3., p. 62), which showed that two-phase study methods did not perform well with very small phase 2 sample sizes due to sparse data, more specifically, very low numbers of people sampled with certain combinations of exposure, outcome and confounder status.

Here we focused on the less common confounder which has a postulated prevalence of 10% in phase 1 data in our simulation settings. We assumed that a proxy measure was available in the phase 1 data with limited sensitivity but excellent specificity. The settings for this simulation are shown in Table 22. They are nearly identical to the first simulation presented in this section (Table 12, Section V.C.1.); the only difference is that now a proxy measure for confounder 1 is available at phase 1, which allows oversampling on the confounder for phase 2 data collection. Parameters that are new in this simulation are shown in bold.

Table 22. Simulation inputs with proxy measure available for confounder 1

Sample size	150,000 users of saxagliptin or sitagliptin
Exposure prevalence	20% saxagliptin, 80% sitagliptin
Prevalence of confounder(s)	10% confounder 1 40% confounder 2
Accuracy of phase 1 proxy measure for confounder 1	**Sensitivity 50%, specificity 99%, Positive predictive value 85%, Negative predictive value 95%**
Strength of confounding (confounder-exposure OR)	OR = 3.00 (confounder 1-saxagliptin) OR = 2.00 (confounder 2-saxagliptin)
Strength of confounding (confounder-outcome OR)	OR = 4.00 (confounder 1-MI) OR = 2.00 (confounder 2-MI)
Expected incidence of the outcome, MI	1/100
Underlying (true) association between exposure and outcome	OR = 1.00 (saxagliptin-MI)

Recall that from the initial simulation, we obtained an estimated OR of 1.44 (95% CI 1.28-1.61) in the unadjusted phase 1 (administrative-data) analysis. Adjustment for the proxy measure of confounder 1 (from administrative data) would be expected to yield a slightly lower OR of 1.31 (95% CI 1.16-1.47) in the phase 1 analysis. Again the true association posited in this simulation is an OR of 1.00.

Table 23 (next page) shows our assumed sampling scheme for phase 2, which includes stratification for the proxy measure of confounder 1.

Compared to a sampling scheme that ignores the proxy measure for confounder 1, sampling based on the proxy measure in addition to outcome and exposure is likely to greatly enrich the phase 2 sample for people with confounder 1. To illustrate, Table 24 (next page) shows how many people with each combination of exposure, outcome, and confounder 1 would be expected to be sampled under a scheme that does not stratify on the proxy measure for confounder 1 vs. a scheme that does. We show expected counts for two different phase 2 sample sizes (250 and 1000). We use bold font to highlight two strata where the difference between the sampling schemes is most striking.

Table 23. Sampling scheme for simulations with different sample sizes for phase 2

				To be sampled at phase 2		
MI	**Saxagliptin Use**	**Confounder 1 (as measured by proxy)**	**N**	**1000**	**500**	**250**
No	No	No	113,347	129	62	31
No	No	Yes	5,548	129	62	31
No	Yes	No	26,539	129	62	31
No	Yes	Yes	3,063	129	62	31
Yes	No	No	965	129	62	31
Yes	No	Yes	142	129	62	31
Yes	Yes	No	299	129	62	31
Yes	Yes	Yes	97	97	62	31

Table 24. Expected number of people selected at phase 2 (for each combination of outcome, exposure, and true confounder 1), comparing a sampling scheme that does not stratify the phase 1 data on a proxy for confounder 1 vs. one that does

				Expected counts in the final phase 2 sample, by presence or absence of stratification on Confounder 1 proxy			
MI	**Saxagliptin Use**	**True Confounder 1**	**N**	**1000, no**	**1000, yes***	**250, no**	**250, yes***
No	No	No	109,999	231	150	57	36
No	No	Yes	8,896	**19**	**108**	**5**	**26**
No	Yes	No	23,954	202	125	50	30
No	Yes	Yes	5,648	**48**	**133**	**12**	**32**
Yes	No	No	841	190	119	47	29
Yes	No	Yes	266	60	139	15	33
Yes	Yes	No	206	130	90	32	22
Yes	Yes	Yes	190	120	135	30	40

*The reason that numbers are not equal in all cells within each of these columns is that the phase 1 sample is stratified on the proxy measure for confounder 1, an imperfect measure, rather than on the true value (not known in phase 1).

Compared to the initial simulation (Sections V.C.2. and 3.), performance of the two-phase design changed little with regard to bias when we stratified and oversampled on a confounder. For phase 2 sample sizes of 250 to 1000, there was very little bias in the OR with either study design. For instance,

with a sample size of 250 in phase 2, with no stratification on a confounder (the initial simulation setting) the mean OR across all simulated trials was 0.95, while with stratification it was 0.96.

Precision of estimates did change somewhat with stratification on the proxy confounder, resulting in narrower CIs. For instance, with a phase 2 sample size of 250, the average width of CIs across all simulation trials was 0.87 on the OR scale with no stratification on the confounder and 0.74 with stratification. It is important to note, though, that the greater precision in the design with stratification on a proxy confounder led to the coverage of the 95% CIs being worse (lower than the proper coverage), suggesting that the CIs were overly narrow. This was seen across all phase 2 sample sizes. For example, with a phase 2 sample of 1000, the initial simulation would have excluded the null (falsely) in 5% of trials, while with stratification on a confounder, 7% of trials falsely excluded the null. With a phase 2 sample size of 250, these values were 8% and 12%, respectively. Thus both versions of the simulation have improper coverage with a phase 2 sample size of 250, but the problem is worse when the sampling design stratifies on the proxy confounder. A benefit, though, was that stratification on the proxy confounder at phase 1 eliminated the issue of "sparse" samples leading to extreme OR estimates (Section V.C.3, pp. 62-63).

In summary, in this particular setting, the two-phase design did not perform substantially better when sampling for phase 2 was stratified on a relatively rare proxy confounder at phase 1 (in addition to stratification by outcome and exposure). Bias changed little with the oversampling. Precision improved, leading to slightly narrower CIs but also to somewhat worse coverage of the 95% confidence intervals (undercoverage).

D. SUMMARY

Our goal in this section was to illustrate how simulations can help guide design choices when the team leading a surveillance activity is determining whether to conduct a two-phase study and if so, how to design that study. We based our simulations on the Mini-Sentinel surveillance activity examining saxagliptin and risk of MI (an activity that has not yet generated a signal requiring investigation). We explored the expected bias, precision, and coverage probability for odds ratio estimation from a two-phase study under a variety of assumptions, but for illustrative purposes, we limited our examples to a logistic regression setting with two-phase estimation performed via a weighted likelihood based approach. It is important to note that this is by no means an exhaustive simulation study; therefore, the results shown above should be viewed simply as an illustration of how simulations can be used as a tool for planning a two-phase study.

With that caveat, we note that with the assumptions that we made and the methodology employed,

- A phase 2 sample size of 500 or greater would be expected to perform well, while a sample size of 100 or lower would perform poorly. A sample size of 250 would perform moderately, with some important limitations;
- Performance was better when there was a true outcome-exposure association (OR 1.5) and worse when the association was null (OR 1.0);
- Performance was worse when stronger confounding was present;
- When the outcome was more rare, precision decreased, but performance was still reasonably good in terms of bias reduction and appropriate coverage of confidence intervals; and

- Performance did not improve overall when we oversampled on a relatively rare confounder for phase 2 (based on a proxy measure).

As noted above, it is important not to attempt to generalize the findings from these examples to all settings. For instance, some additional simulations not available at the time of the writing of this report suggest that perhaps using profile likelihood rather than weighted likelihood methods would substantially improve performance characteristics in many of the settings considered above (especially the small sample settings). However, those improvements may come at a cost, as profile likelihood methods may be strongly affected when there is model misspecification. To reach more generalizable conclusions, a broader and richer formal simulation study would be needed, one that more fully evaluates a range of simulation input parameters and estimation methodologies. Still, Section V illustrates that simulation studies can be helpful to guide study design choices when a team wishes to carry out a two-phase study as part of determining whether a potential safety signal is likely to be valid.

E. RESOURCES TO SUPPORT SIMULATIONS IN FUTURE MINI-SENTINEL SURVEILLANCE ACTIVITIES

For this workgroup, we developed a program to carry out variations on the types of simulation studies presented above. We are making this program available to Mini-Sentinel for future use. This program is written in R with some basic documentation explaining how to use it. These materials will be available through the Mini-Sentinel Operations Center or by writing directly to Dr. Dublin (dublin.s@ghc.org) and Mr. Walker (walker.rl@ghc.org) at Group Health Research Institute.

Figure 4: Histograms of estimated odds ratios (OR) for the association between outcome (Y) and exposure (X) across simulation trials for 4 different phase 2 sampling designs. Panels (from left to right) show results based on phase 2 sample sizes of 1000, 500, 250, and 100, respectively.

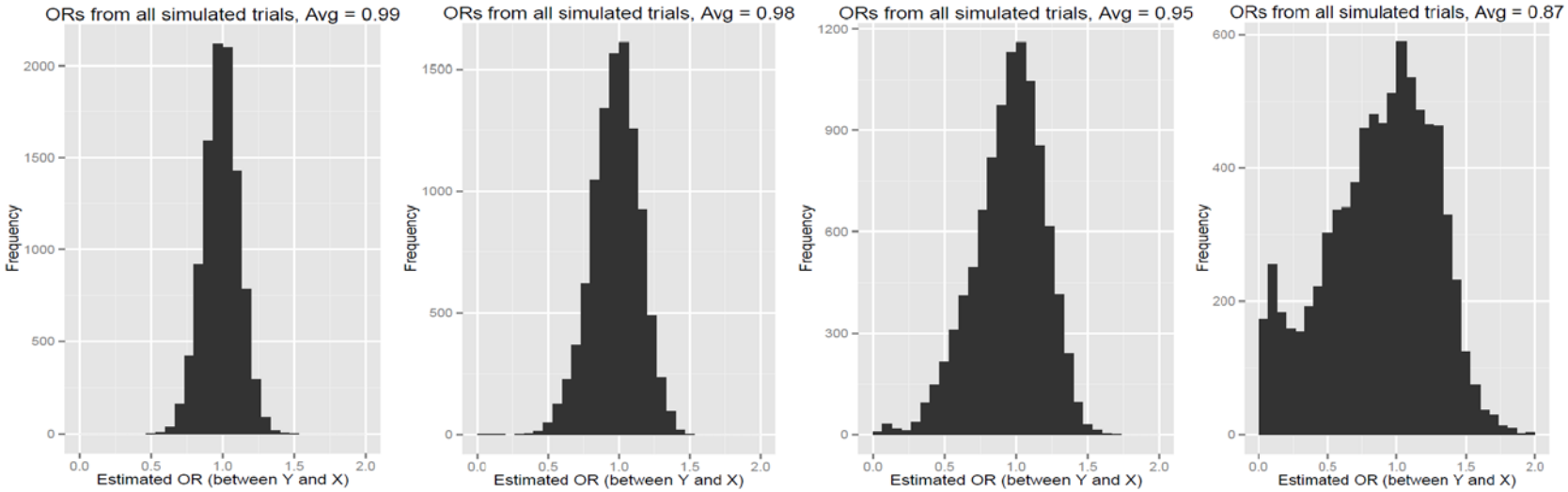


Figure 5: Examples of estimated odds ratios (OR) and 95% confidence intervals (CI) for the association between outcome (Y) and exposure (X) across simulation trials for 4 different phase 2 sampling designs. Panels (from left to right) show results based on phase 2 sample sizes of 1000, 500, 250, and 100, respectively.

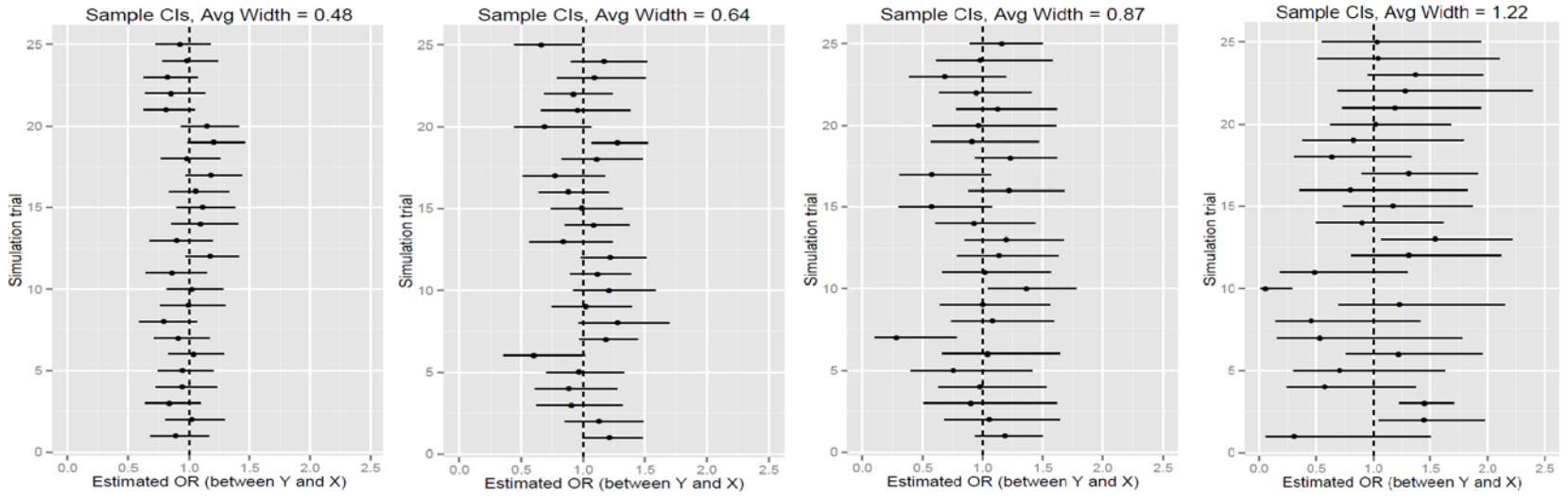
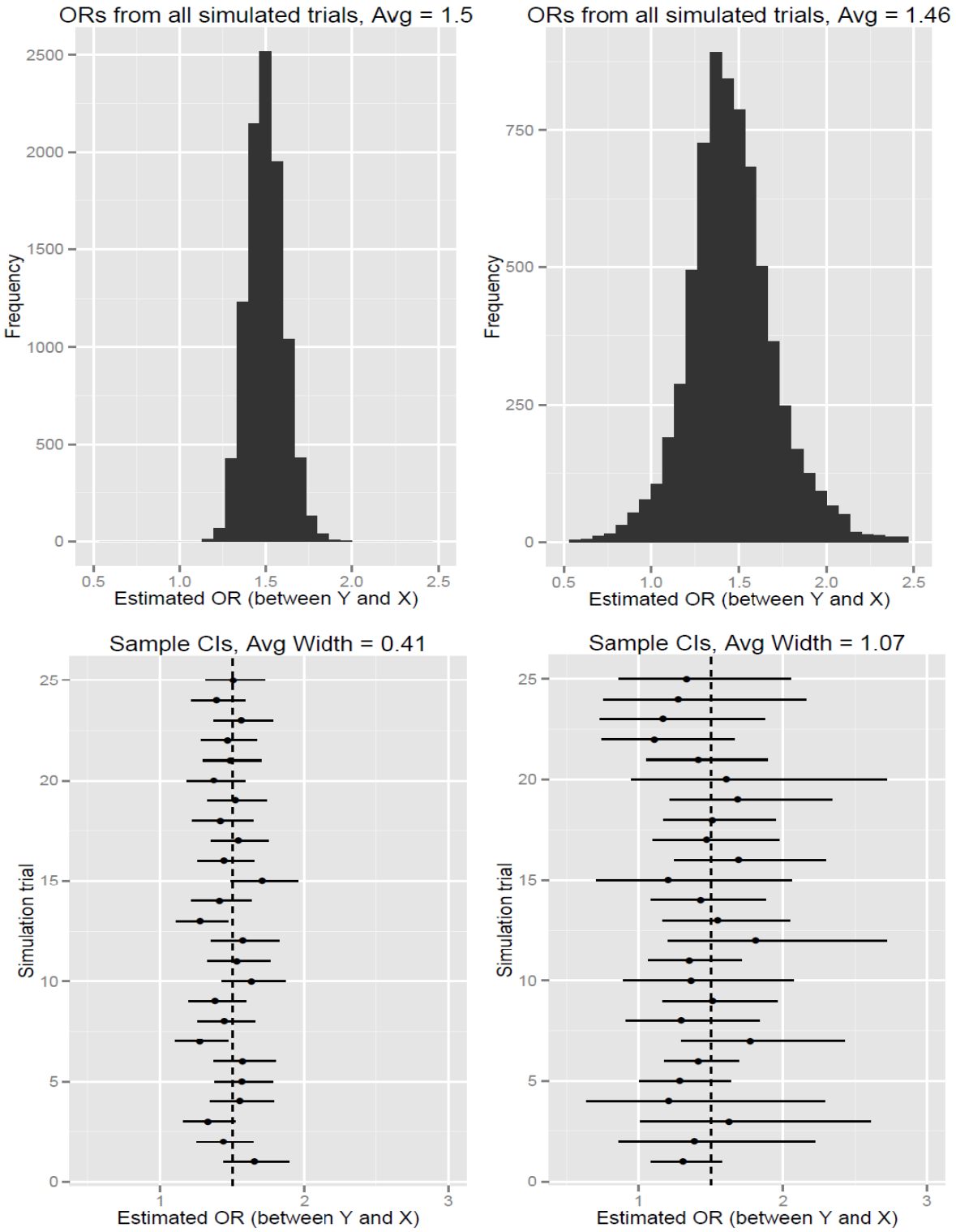


Figure 6: Simulation results showing estimated odds ratios (OR) and 95% confidence intervals (CI) for the association between outcome (Y) and exposure (X) across simulation trials for 2 different phase 2 sampling designs. Panels on the left are for a phase 2 sample size of 1000; panels on the right are for a phase 2 sample size of 100.



VI. REFERENCES

1. McClure DL, Raebel MA, Yih WK, Shoaibi A, Mullersman J, Anderson-Smits C, Ouellet-Hellstrom R, Chakravarty A, Kim C, Glanz J. Framework for Assessment of Signal Refinement Positive Results. Final report prepared for Mini-Sentinel. 2012 [cited Oct 31 2013]; Available from: http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Framework-for-Assessment-of-Signal-Refinement-Positive-Results.pdf.
2. McClure DL, Raebel MA, Yih WK, Shoaibi A, Mullersman J, Anderson-Smits C, Ouellet-Hellstrom R, Chakravarty A, Kim C, Glanz J. Mini-Sentinel methods: Framework for assessment of signal refinement positive results. Pharmacoepidemiol Drug Saf. 2013(in press).
3. Collet JP, Schaubel D, Hanley J, Sharpe C, Boivin JF. Controlling confounding when studying large pharmacoepidemiologic databases: A case study of the two-stage sampling design. Epidemiology. 1998;9(3):309-15.
4. McNamee R. Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures. Biostatistics. 2005;6(4):590-603.
5. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988;75(1):11-20.
6. Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. Stat Med. 1997;16(1-3):103-16.
7. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. Appl Stat. 1999;48:457-68.
8. Schaubel D, Hanley J, Collet JP, Bolvin JF, Sharpe C, Morrison HI, Mao Y. Two-stage sampling for etiologic studies. Sample size and power. Am J Epidemiol. 1997;146(5):450-8.
9. Hanley JA, Dendukuri N. Efficient sampling approaches to address confounding in database studies. Stat Methods in Med Research. 2009;18:81-105.
10. Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, Ng D, Butler MG, Boudreau D, Forrow S, Goldberg R, Gore J, McManus D, Racoosin JA, Gurwitz JH. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf. 2013;22(1):40-54. PMCID: PMC3601831.
11. Lo Re V. III, Haynes K, Goldberg D, Forde KA, Carbonari DM, Leidl KB, Hennessy S, Reddy KR, Pawloski PA, Daniel GW, Cheetham TC, Iyer A, Coughlin KO, Toh S, Boudreau DM, Selvam N, Cooper WO, Selvan MS, Vanwormer JJ, Avigan MI, Houstoun M, Zornberg GL, Racoosin JA, Shoaibi A. Validity of diagnostic codes to identify cases of severe acute liver injury in the U.S. Food and Drug Administration's Mini-Sentinel Distributed Database. Pharmacoepidemiol Drug Saf. 2013;22(8):861-72.
12. Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet. 1955;19(4):251-3.
13. Haneuse S, Schildcrout J, Gillen D. A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. Biostatistics. 2012;13(2):274-88. PMCID: PMC3297823.
14. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J R Stat Soc Series B Stat Methodol. 1997;59(2):447-61.
15. Haneuse S, Saegusa T, Lumley T. osDesign: An R Package for the Analysis, Evaluation, and Design of Two-Phase and Case-Control Studies. J Stat Softw. 2011;43(11).

16. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Stat Med. 1991;10(5):739-47.
17. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. JASA. 1952;47(260):663-85.
18. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol. 1982;115(1):119-28.
19. Huber PJ. The behavior of maximum likelihood estimated under nonstandard conditions. In: Neyman, editor. Proceedings of the 5th Berkley Symposium in Mathematical and Statistical Probability. Berkeley, CA 1967. p. 221-33.
20. Schill W, Jockel KH, Drescher K, Timm J. Logistic analysis of case-control studies under validation sampling. Biometrika. 1993;80:339-52.
21. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997;84(1):57-71.
22. Lee AJ, Scott AJ, Wild CJ. Efficient estimation in multi-phase case-control studies. Biometrika. 2010;97(2):361-74.
23. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997;84(1):57-71.
24. Scott AJ, Wild CJ. Fitting logistic regression models in stratified case-control studies. Biometrics. 1991;47(2):497-510.
25. Lumley T, Scott A. Partial likelihood ratio tests for the Cox model under complex sampling. Stat Med. 2013;32(1):110-23.
26. Breslow NE, Wellner JA. A Z-theorem with Estimated Nuisance Parameters and Correction Note for 'Weighted Likelihood for Semiparametric Models and Two-phase Stratified Samples, with Application to Cox Regression'. Scand Stat Theory Appl. 2008;35(1):186-92.
27. Breslow NE, Lumley T. Semiparametric models and two-phase samples: applications to Cox regression. In: Banerjee M, Bunea F, Huang J, Koltchinskii, Maathuis MH, editors. Probability to Statistics and Back: High-Dimensional Models and Processes Essays in Honor of Jon Wellner's 65th Birthday: IMS Collections; 2013. p. 65-77.
28. Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. Scand J Stat. 2007;34(1):86-102.
29. Therneau TM, Grambsch. Modeling survival data: Extending the cox model. New York: Springer; 2000.
30. Ross M, Wakefield J. Bayesian inference for two-phase studies with categorical covariates. Biometrics. 2013;69(2):469-77.
31. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol. 2009;169(11):1398-405. PMCID: PMC2768499.
32. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz-Thompson Estimation of Model Parameters from Two-phase Stratified Samples: Applications in Epidemiology. Stat Biosci. 2009;1(1):32.
33. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005;162(3):279-89.
34. Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London: Chapman & Hall; 1995.
35. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration--a simulation study. Am J Epidemiol. 2007;165(10):1110-8. PMCID: PMC1945235.

36. Lunt M, Glynn RJ, Rothman KJ, Avorn J, Sturmer T. Propensity score calibration in the absence of surrogacy. Am J Epidemiol. 2012;175(12):1294-302. PMCID: PMC3491974.
37. Fireman B, Toh S, Butler MG, Go AS, Joffe HV, Graham DJ, Nelson JC, Daniel GW, Selby JV. 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.
38. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150(4):327-33.
39. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
40. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006;98(3):253-9. PMCID: PMC1790968.
41. Cadarette SM, Gagne JJ, Solomon DH, Katz JN, Sturmer T. Confounder summary scores when comparing the effects of multiple drug exposures. Pharmacoepidemiol Drug Saf. 2010;19(1):2-9. PMCID: PMC2800174.
42. Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. Lancet. 2002;360(9339):1071-3.
43. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Stat Methods Med Res. 2009;18(1):67-80.
44. Psaty BM, Koepsell TD, Lin D, Weiss NS, Siscovick DS, Rosendaal FR, Pahor M, Furberg CD. Assessment and control for confounding by indication in observational studies. J Am Geriatr Soc. 1999;47(6):749-54.
45. Jackson LA, Nelson JC, Benson P, Neuzil KM, Reid RJ, Psaty BM, Heckbert SR, Larson EB, Weiss NS. Functional status is a confounder of the association of influenza vaccine and risk of all cause mortality in seniors. Int J Epidemiol. 2006;35(2):345-52.
46. Jackson ML, Nelson JC, Weiss NS, Neuzil KM, Barlow W, Jackson LA. Influenza vaccination and risk of community-acquired pneumonia in immunocompetent elderly people: a population-based, nested case-control study. Lancet. 2008;372(9636):398-405.
47. Majumdar SR, McAlister FA, Eurich DT, Padwal RS, Marrie TJ. Statins and outcomes in patients admitted to hospital with community acquired pneumonia: population based prospective cohort study. BMJ (Clinical research ed). 2006;333(7576). PMCID: PMC1625620.
48. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology. 2001;12(6):682-9.
49. Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. Int J Epidemiol. 2006;35(2):337-44.
50. Ohlsson H, Chaix B, Merlo J. Therapeutic traditions, patient socioeconomic characteristics and physicians' early new drug prescribing--a multilevel analysis of rosuvastatin prescription in south Sweden. Eur J Clin Pharmacol. 2009;65(2):141-50.
51. Hlatky MA, Cotugno H, O'Connor C, Mark DB, Pryor DB, Califf RM. Adoption of thrombolytic therapy in the management of acute myocardial infarction. Am J Cardiol. 1988;61(8):510-4.
52. Hirth RA, Fendrick AM, Chernew ME. Specialist and generalist physicians' adoption of antibiotic therapy to eradicate Helicobacter pylori infection. Med Care. 1996;34(12):1199-204.
53. Lee DS, Stitt A, Wang X, Yu JS, Gurevich Y, Kingsbury KJ, Austin PC, Tu JV. Administrative hospitalization database validation of cardiac procedure codes. Med Care. 2013;51(4):e22-6.
54. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care. 2005;43(5):480-5.

55. Kokotailo RA, Hill MD. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. Stroke. 2005;36(8):1776-81.
56. Walsh CO, Milliren CE, Feldman HA, Taveras EM. Sensitivity and Specificity of Obesity Diagnosis in Pediatric Ambulatory Care in the United States. Clin Pediatr (Phila). 2013.
57. Schneeweiss S, Wang PS. Association between SSRI use and hip fractures and the effect of residual confounding bias in claims database studies. J Clin Psychopharmacol. 2004;24(6):632-8.
58. Floyd JS, Heckbert SR, Weiss NS, Carrell DS, Psaty BM. Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. JAMA. 2012;307(15):1580-2.
59. Boger-Megiddo I, Heckbert SR, Weiss NS, McKnight B, Furberg CD, Wiggins KL, Delaney JA, Siscovick DS, Larson EB, Lemaitre RN, Smith NL, Rice KM, Glazer NL, Psaty BM. Myocardial infarction and stroke associated with diuretic based two drug antihypertensive regimens: population based case-control study. BMJ. 2010;340:c103. PMCID: PMC2811239.
60. Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. J Am Med Inform Assoc. 2013;20(4):652-8.
61. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Identifying hypertension-related comorbidities from administrative data: what's the optimal approach? Am J Med Qual. 2004;19(5):201-6.
62. Heckbert SR, Kooperberg C, Safford MM, Psaty BM, Hsia J, McTiernan A, Gaziano JM, Frishman WH, Curb JD. Comparison of self-report, hospital discharge codes, and adjudication of cardiovascular events in the Women's Health Initiative. Am J Epidemiol. 2004;160(12):1152-8.
63. Newton KM, Wagner EH, Ramsey SD, McCulloch D, Evans R, Sandhu N, Davis C. The use of automated data to identify complications and comorbidities of diabetes: a validation study. J Clin Epidemiol. 1999;52(3):199-207.
64. Andrade SE, Harrold LR, Tjia J, Cutrona SL, Saczynski JS, Dodd KS, Goldberg RJ, Gurwitz JH. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1:100-28. PMCID: PMC3412674.
65. Quan H, Parsons GA, Ghali WA. Validity of information on comorbidity derived from ICD-9-CCM administrative data. Med Care. 2002;40(8):675-85.
66. Jackson ML, Nelson JC, Jackson LA. Why do covariates defined by International Classification of Diseases codes fail to remove confounding in pharmacoepidemiologic studies among seniors? Pharmacoepidemiol Drug Saf. 2011;20(8):858-65.
67. Koroukian SM, Xu F, Murray P. Ability of Medicare claims data to identify nursing home patients: a validation study. Med Care. 2008;46(11):1184-7. PMCID: PMC3178883.
68. Zuckerman IH, Sato M, Hsu VD, Hernandez JJ. Validation of a method for identifying nursing home admissions using administrative claims. BMC Health Serv Res. 2007;7:202. PMCID: PMC2222626.