# MINI-SENTINEL METHODS

# ANALYTIC METHODS FOR USING LABORATORY TEST RESULTS IN ACTIVE DATABASE SURVEILLANCE: FINAL REPORT

**Prepared by:** Marsha A. Raebel, PharmD,[1] Susan Shetterly, MS,[1] Andrea R. Paolino, MA,[1] Christine Y. Lu, PhD, MSc,[2] Joshua J. Gagne, PharmD, ScD,[3] Kevin Haynes, PharmD, MSCE,[4] James Flory, MD, MS,[5] Elisabetta Patorno, MD, DrPH,[3] David H. Smith, PhD,[6] Mano Selvan, PhD,[7] Lisa J. Herrinton, PhD,[8]Frank E. Harrell Jr., PhD,[9] Azadeh Shoaibi, PhD,MHS,[10] Jason Roy, PhD,[11]

**Author Affiliations:** [1]Kaiser Permanente Colorado, Denver, CO, [2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, [3]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, [4]HealthCore, Wilmington, DE, [5]Department of Medicine, Weill Cornell Medical College, New York, NY, [6]Kaiser Permanente Northwest, Portland, OR, [7]Comprehensive Health Insights, Humana Inc., Louisville, KY, [8]Kaiser Permanente Northern California, Oakland, CA, [9]Department of Biostatistics, Vanderbilt University School of Medicine Nashville, TN, [10]Center for Biologics Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD, [11]University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

**May 16, 2016**

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to healthcare 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.

## Acknowledgements

**Additional Participating Members of the Analytic Methods for Using Laboratory Test Results in Active Database Surveillance Workgroup**

| Affiliation | Name |
|---|---|
| Kaiser Permanente Colorado | Carsie Nyirenda, MPH |
| | Andrew T. Sterrett, PhD |
| University of Pennsylvania | Xingmei Wang, MS |
| Center for Drug Evaluation and Research, United States Food and Drug Administration | Michael Nguyen, MD |
| | Carlos Bell, MPH |
| | Aaron Niman, MPH |
| | Ling Lan, PhD |
| | Clara Kim, PhD |
| Harvard Pilgrim Health Care Institute | Susan Forrow, MS |
| | Madelyn Pimentel, BA |
| | Jennifer Popovic, DVM, MA |
| | Lisa Trebino Ortendahl, MS |

# Mini-Sentinel Methods

# Analytic Methods for Using Laboratory Test Results in Active Database Surveillance: Final Report

## Table of Contents

# I. EXECUTIVE SUMMARY

Clinical data such as laboratory test results are obtained as part of routine healthcare delivery. These data are not collected primarily for research and are not always available or complete. If laboratory results are included in analyses without considering the missing data, estimates can be biased. To appropriately use laboratory results data, robust statistical tools can be required to manage missing data.

The objectives of this project included selecting and testing statistical methods for use when administrative claims and electronic health records (EHR) data are enriched with clinical laboratory test results, and the laboratory test results have missing data. Three applications were included in this project, including baseline confounding adjustment, cohort identification, and outcomes detection. The Specific Aims were to: 1) Summarize the literature on use of clinical laboratory results in administrative claims and EHR database studies of medical product safety; 2) Utilize the Mini-Sentinel Distributed Database (MSDD) resources to evaluate statistical approaches to incorporate clinical laboratory results data into medical product safety analyses; and 3) Develop recommendations and specifications to incorporate laboratory results data into Mini-Sentinel safety analyses.

***Findings from Specific Aim 1: Several sources of missing laboratory results data exist in the MSDD. However, the published literature utilizing clinical laboratory results focused almost entirely on Patient-Level missingness*** (patient characteristics or non-adherence or selective test ordering by providers). ***The literature is of little assistance in informing handling of Facility-, Organization-, Care Setting-, or Temporal-Level sources of missing data.***

Mechanisms of missing data include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR rarely holds for laboratory results data and statistical approaches that assume MCAR, such as complete case analysis, are not appropriate to use when analyzing laboratory results data. MAR mechanisms could occur with laboratory results, but whether data are MAR is difficult to confirm. MNAR may also occur with laboratory results, but few publications employed missing data techniques to account for MNAR data. ***Consistent missing data mechanisms should not be expected with laboratory results data.*** The missing data mechanisms in a given dataset are influenced by the laboratory tests of interest, the population characteristics, and provider and system factors.

In the literature, use of missing data techniques when analyzing laboratory results data was relatively uncommon. Further, few publications included data from multiple sites and different organization types, which is the usual condition in Mini-Sentinel. Across available publications, ***multiple imputation (MI) methods emerged as important. The relative prominence of MI was likely influenced by both the depth of prior methodological work as well as increasing availability of MI methods in software packages.***

***Findings from Specific Aim 2, laboratory results as baseline confounder:*** Three baseline confounder test cases were studied: 1) Baseline glycosylated hemoglobin (HbA1c) or fasting blood glucose or random blood glucose (composite variable "GLU") in a cohort of adults without diabetes newly-initiating a second generation antipsychotic (SGA) where the outcome of interest was a diabetes diagnosis; 2) Baseline serum creatinine in a cohort of patients with diabetes who were newly-starting lisinopril where the outcome of interest was a hyperkalemia diagnosis; and 3) Baseline INR in a cohort of current warfarin users starting an antimicrobial medication where the outcome of interest was a bleeding diagnosis. In all three test cases, the amount of missing laboratory results data varied across the three

participating data partner sites, with the highest amount of missing data at the large national insurer data partner. Important site differences were observed in both predictors of missingness and in outcomes. The most striking finding from these test cases was the differences in results between a single pooled analysis versus either a site-specific analysis or a meta-analytic approach. ***The workgroup therefore recommends site-specific models when using laboratory results as baseline confounders.*** Several different missing data methods and models led to similar point estimates and 95% confidence intervals in all baseline confounder test cases. Models that ignored missing data at times yielded different results from models that included missing data. ***Different analytic methods, particularly the approach taken to integrate results from different Data Partner sites, impacted model results more than different missing data methods.***

***Findings from Specific Aim 2, laboratory results to supplement cohort identification:*** Two cohort identification test cases were considered: 1) Supplementing identification of a cohort of pregnant women through considering positive qualitative or high quantitative pregnancy laboratory results (human chorionic gonadotropin; HCG) in addition to prenatal diagnosis and procedure codes; and 2) Enhanced identification of a cohort of adults with Chronic Kidney Disease (CKD) through considering serum creatinine laboratory results to estimate patients' glomerular filtration rates in addition to coded CKD diagnosis. ***Important variability in missing laboratory results was again observed across Data Partners. In both test cases, inclusion of laboratory results augmented the cohort size at all Data Partner sites.*** Using laboratory results identified individuals with CKD who would not have been identified using definitions that did not include laboratory values. Also, in both cohorts, some individuals were recognized for cohort inclusion earlier with laboratory results than with diagnoses. Use of HCG results was particularly important when identifying women with miscarriages and abortions. For studying drug safety in pregnancy, identifying pregnancies based on laboratory results might be important if the exposure drug could be associated with maternal safety concerns, pregnancy loss, or other adverse impacts on the fetus or infant.

***Findings from Specific Aim 2, laboratory results to detect outcomes:*** Two outcomes detection test cases were considered: 1) Diabetes outcome and blood glucose or HbA1c laboratory test results (GLU) among adults initiating an SGA; and 2) Upper gastrointestinal (UGI) bleeding outcome and hemoglobin (HGB) laboratory test results among adults initiating a non-steroidal anti-inflammatory drug (NSAID). Once again, ***missing laboratory results data varied across data partner sites. In the first outcomes test case, laboratory results identified additional patients with the outcome, but including laboratory results did not identify outcomes earlier. The outcome rate was higher among individuals with laboratory results available, suggesting that clinicians selectively choose patients for laboratory monitoring who they believe are at higher risk of the outcome. This finding could also suggest that outcomes are under-recognized due to lack of monitoring.*** In the second outcomes test case, inpatient UGI bleeding diagnosis codes (with or without an observed drop in HGB $\geq$ 3 g/dL) was considered the standard definition. ***Using an observed drop in HGB in conjunction with a non-inpatient coded UGI bleeding diagnosis identified very few additional outcomes across the three participating sites. Further, using HGB results alone to identify UGI bleeding did not distinguish between UGI bleeding and bleeding from other body locations.*** Therefore, a drop in HGB was only modestly useful in detecting clinically important bleeding, and a drop in HGB alone was not sufficient to identify UGI bleeding.

***Findings from Specific Aim 3:*** The literature is of modest value in guiding approaches to handling missing laboratory results data in Mini-Sentinel assessments. ***When developing an active surveillance plan, the missing data should be described overall and by key study population variables early in the project.***

***Reasons for missing laboratory results in observational multi-site studies are complex and deserve thought and discussion during analysis planning.***

***Including laboratory results data can improve baseline confounding control, with a few caveats.*** For baseline confounder assessment using MSDD data, site-specific models are important for MI and propensity scores even when end results are pooled across sites. The MI regression method is common and easy to implement. Because it relies on a normality assumption, skewness must be assessed up front and a log-transformation should be performed to improve symmetry if needed. Ideally, results should be compared between different strategies for handling laboratory results data. ***Including laboratory results as baseline confounders using statistical tools such as MI is a decision that should be made after considering the amount of missing data at the data partner sites. Including laboratory results data should not be based primarily on imputed data from some Data Partners.***

***We recommend including laboratory results data (when applicable) to supplement cohort identification when applying an algorithmic approach.*** Comparing percentages of the cohort from each site identified using laboratory results data is an important component of assessing appropriateness within specific cohorts. Utilizing the laboratory results can be particularly important to drug safety studies, as we demonstrated in the pregnancy cohort identification and the CKD cohort identification test cases we examined.

***We do not recommend employing laboratory results data as a sole criterion when detecting health outcomes. In general, we also do not recommend imputing outcomes.*** In one of the outcomes test cases, laboratory results identified an important number of additional outcomes, while in the other, few additional outcomes were identified. ***Decisions about whether or not to include laboratory results in algorithms that identify health outcomes in the MSDD should be made on a scenario-by-scenario basis.***

Although not part of the scope of work of this project, it was intended that findings from this Workgroup would ultimately lead to the development of software programs for incorporating laboratory results into safety analyses that are compatible with existing modular programs for expedited safety monitoring using the Mini-Sentinel Routine Analytic Framework. ***The existing Mini-Sentinel Routine Analytic Framework can identify some cohorts that utilize laboratory results data.*** However, the existing Mini-Sentinel Routine Analytic Framework is not sufficient for utilizing laboratory results data in scenarios that have complex index dates, that require temporal relationships between laboratory results, or that utilize changes in laboratory result values over time. Enhancements to existing capabilities would be necessary to enable use of laboratory results in those scenarios. Also, the Mini-Sentinel Routine Analytic Framework currently cannot assist with missing data characterization. Logical first steps to develop such capability could be creating a repeatable, reusable process to describe missing data within a cohort overall and by site, and diagnosing whether available laboratory test result data appear to be MAR to inform whether it is reasonable to impute the missing data. Other tasks required to use laboratory results currently require custom programming such as the capability to then incorporate missing data techniques when missing data are expected. ***Given existing capabilities of the Mini-Sentinel Routine Analytic Framework, use of laboratory results data in current Mini-Sentinel evaluations is most often feasible only within Protocol-Based Assessments.***

# II. INTRODUCTION

As part of the activities for Task Order #8 (Foundational Elements 2 /Medical Countermeasures, Activity E: Statistical Methods Development) several statistical methods were selected and developed to improve Mini-Sentinel active surveillance capabilities. One of these activities was to develop and test analytic methods for using laboratory test results in active surveillance for improving confounding control and outcome definitions. By 2014, the Mini-Sentinel Distributed Database (MSDD) clinical laboratory results data table (LRT) included over 730 million clinical laboratory test results for selected chemistry, hematology, coagulation and influenza tests obtained from nearly 32 million unique health records.[1] Appropriate use of these data is not straightforward because laboratory results data carry availability and completeness issues reflecting several missing data mechanisms. Inclusion of LRT data in active surveillance analyses without considering the missing data or even with applying older analytic techniques based on fairly strict assumptions about the causes of missing data can yield markedly biased estimates.[2] Some recent approaches to handling missing data (e.g., multiple imputation [MI]) are potentially useful because they require less strict assumptions about the causes of missing data, and can produce less biased parameter estimates.[2] However, even newer statistical techniques often require data to be missing at random (MAR). Therefore, testing the performance of statistical approaches and developing guidance for appropriately including LRT in Mini-Sentinel surveillance is critical.

The objective of this project was to select and test statistical methods appropriate for use when analyzing cross-sectional and longitudinal observational healthcare administrative, claims, and clinical data, with specific attention to clinical laboratory results when missing data are expected. Recognizing that different uses of laboratory results data (e.g., confounding adjustment, detecting outcomes, cohort identification) in medical product safety surveillance could optimally employ different methods to handle the missing results data, we evaluated more than one method for use in the MSDD environment.

To select and evaluate the performance of statistical methods with intent to potentially recommend techniques for including laboratory test results data in Mini-Sentinel safety surveillance, we had the following Specific Aims:

- Specific Aim 1. Summarize the literature on the use of clinical laboratory test results databases in administrative claims and electronic health records (EHR) database studies of medical product safety.
- Specific Aim 2. Utilize the MSDD resources to evaluate analytic methods for incorporating laboratory test results data into medical product safety analyses.
- Specific Aim 3. Develop recommendations and detailed specifications for incorporating LRT data into Mini-Sentinel safety analyses.

# III. MISSING DATA IN THE MINI-SENTINEL COMMON DATA MODEL (MSCDM) LABORATORY RESULTS DATA TABLE (LRT)

## A. SOURCES OF MISSING DATA IN THE LRT

The MSDD (www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/106) is emerging as an efficient tool for active surveillance of medical product safety.[3] Data transformed into the MSCDM at each Data Partner (DP) site facilitate use of shared programming across DP sites, while maintaining local control, security, and confidentiality of data access and use.[4-6] The MSCDM includes linked tables comprised of transformed administrative and claims data for enrollment, enrollee demographics, diagnoses, encounters, procedures, outpatient pharmacy dispensings, and death.[7, 8] While claims for laboratory procedures are in the MSCDM tables and indicate laboratory tests were completed, procedure claims do not include clinical result values and are of little use for confounder adjustment, cohort identification, or outcomes detection associated with safety of a medical product exposure. However, laboratory test result values have the potential to contribute to confounder adjustment, enhance cohort identification, and supplement outcomes definitions, and availability of electronically-extractable laboratory results at several DP sites made it possible to expand the MSCDM to include an LRT.

Because LRT data are observational and extracted from data collected during routine healthcare delivery at US medical facilities, they carry availability and completeness challenges. There are two fundamental contributors to missingness of laboratory results data in the MSCDM. First, the laboratory test was not done because it was never ordered or because an ordered test was not successfully completed. Second, the laboratory test was completed but the result was not accessible for inclusion in the MSCDM due to administrative, contracting, database, or documentation factors. Therefore, to appropriately use LRT data in medical product safety surveillance requires applying statistical tools that aid in managing and minimizing these inherent missing data challenges. The LRT data availability or completeness challenges are detailed below.

### 1. Differential missing data across Data Partners (DPs) and variation in laboratory test result data availability by DP type (DP Organization-Level Missingness)

Currently, 12 (of 18) DPs contribute laboratory results data to the MSDD, including three large insurance plans, three healthcare delivery systems with insurance functions, and six group model integrated healthcare delivery systems. These 12 DPs accounted for about 100 of the 150+ million unique individual health records in the MSDD in 2013, but not all of these individuals would have laboratory test results (e.g., healthy children are unlikely to have any completed laboratory results). Eleven DPs provide results for all laboratory test types currently in the LRT. One DP provides results for only some laboratory test types. Additional DPs may in the future contribute laboratory results data to the LRT. The LRT data currently available from the three large insurance plans are limited to outpatient tests. Further, these large insurers only have contracts with laboratory systems database vendors that provide results for 15% to 30% of their enrollees. DPs that are healthcare delivery systems with insurance functions have internal laboratory system databases, electronic health record (EHR) databases, and/or contracts with laboratory systems database vendors that provide outpatient test results for 60% to 80% of their members. The group model integrated healthcare delivery system DPs have internal laboratory system/EHR databases that can provide outpatient test results for 90% to 100% of their members.

### 2. Patient location where laboratory testing was conducted and lack of availability of hospital, emergency department, and medical office-based laboratory test results (Care Setting- Level Missingness)

As mentioned above, the laboratory results available from the large insurance plans are limited to test results nearly entirely from the ambulatory care environment. Similarly, the results available from the healthcare delivery systems with insurance functions and from two of the six integrated healthcare delivery systems are limited primarily to outpatient results. Four integrated healthcare delivery systems provide outpatient test results and also provide inpatient test results from acute care hospitalizations for 40% to 100% of their members. One integrated healthcare delivery system DP also has emergency department test results for essentially 100% of its members. Active surveillance activities designed to evaluate severe or life-threatening adverse outcomes where the most abnormal laboratory results are obtained in conjunction with an emergency department visit or inpatient hospitalization (e.g., liver function tests associated with acute liver injury) are underrepresented in the current LRT. Active surveillance activities designed to evaluate population-based laboratory test results obtained in the ambulatory care environment (e.g., international normalized ratio [INR] monitoring associated with warfarin use; glycosylated hemoglobin [HbA1c] results among patients with diabetes) will have more complete data availability capture, particularly at DPs with results data from 90% to 100% of members. For all DPs, laboratory results obtained from testing conducted in the medical office ("point-of-care") are often not available to the DP to include in the LRT.

### 3. Earliest date of laboratory results data availability (Temporal Level Missingness)

LRT data are available from 2006 forward for ten (of 12) DPs. The start date for LRT data from one large insurance plan is mid-2007 and for another large insurance plan is 2008.

### 4. Obtaining laboratory test outside of contracted/owned facilities/laboratories (Facility-Level Missingness)

Test results that are obtained at non-contracted facilities are not available in DP source laboratory databases. Examples include test results that are faxed, called, or hand carried from an outside facility to a DP facility, results obtained during an emergency department visit (for 11 of these 12 DPs), and some point-of-care (POC) test results. Such test results are typically entered into the medical record as a text field or are scanned into the medical record as a PDF (i.e., not electronically extractable into the laboratory database).

### 5. Variation in reference ranges and laboratory assay methods across institutions and within institutions over time (Institutional-Level Missingness)

The scope of this project did not include addressing variation in laboratory results data that occur due to variations in reference ranges and laboratory assay methods, but we briefly describe these for completeness. Different assay methodologies have differing standards and sources of variability, but the assay method can only occasionally be identified from source data. For example, when Logical Observation Identifiers Names and Codes (LOINC®) are associated with test results, the assay method can at times be identified from the specific LOINC®.

The reference range associated with each patient-specific test result from the data partner's source data is included as an LRT data field. This enables applying or adjusting for the reference range when its variability is important to a specific analysis. Existing normalization methods intended to support

integration of clinical laboratory test results from different institutions have not been applied to the MSDD LRT to date because such methods need further development before they will be of potential value to LRT data applications.

### 6. Termination of plan membership (Patient-Level Missingness)

This source of missing data is not unique to the LRT or the MSDD, but is pervasive in observational healthcare databases. Throughout the US, an individual's healthcare data are not recorded into a healthcare provider/insurer's database after the date that individual's enrollment ends with that healthcare provider/insurer.

### 7. Provider non-adherence to ordering recommended laboratory monitoring or patient non-adherence to completing ordered tests (Patient-Level Missingness)

This type of missing data is also pervasive in healthcare databases across the US. Numerous factors contribute to non-adherence with ordering and completing recommended laboratory monitoring some of which include communication gaps between provider and patient, low patient health literacy, lack of knowledge by the provider of recommended monitoring for the drug, and provider perception that laboratory monitoring for a particular drug is not a priority. One example of the extents of provider and patient non-adherence to ordering and completing recommended laboratory monitoring is demonstrated in a randomized controlled trial conducted by Raebel et al.[9] That trial was designed to determine whether a computerized tool that alerted pharmacists to missing laboratory test results was effective at increasing the percentage of patients receiving appropriate monitoring at initiation of high-risk drug therapy.[10] In collaboration with physicians, pharmacists were alerted to missing test results, ordered missing tests, reminded patients to obtain tests, assessed test completion, and reviewed and managed abnormal results. Results included that patients in the intervention group received 5153 dispensings of high-risk drugs within the one year study period versus 5016 dispensings in the control group. For 816 intervention group patients, pharmacists ordered laboratory testing that was recommended at initiation of therapy and that had not been ordered by the physician. For 194 intervention group patients, physicians had ordered laboratory tests that patients had not completed and pharmacists reminded the patients to obtain the tests.

### 8. Selective assessment correlated with patient clinical features, quality initiatives, or care coordination (Patient-Level Missingness)

This type of missing data is also common in observational healthcare databases across the US. Missingness associated with selective assessment can be thought of as confounding by indication due to factors such as the patient's engagement (or non-engagement) with medical care, disease severity, acute illness assessment, chronic illness management, new medication initiation, or other situations where laboratory tests are carried out more frequently in one subset of individuals than another. Also, laboratory test ordering and completion is conducted more often for patients and laboratory test types included in local or national quality measures or in quality improvement initiatives (e.g., measurement of HbA1c within specified time intervals among patients with diabetes is a Health Effectiveness Data and Information Set [HEDIS] quality of care measure). Finally, variations in practice and care coordination across sites and types of healthcare systems contribute to differences in rates of laboratory testing and therefore differences in patient-level missingness of laboratory test results. General examples at several integrated healthcare delivery system sites that participate in Mini-Sentinel are the collaborative drug therapy management protocols and services that facilitate medication management and laboratory

monitoring of medication therapy by a variety of healthcare professionals for chronic conditions such as hypertension, diabetes, anemia, and lipid disorders.[11-13] One specific example is the centralized anticoagulation management service at one of the Mini-Sentinel Data Partners that, in comparison to usual care, has been shown to increase the frequency of international normalized ratio (INR) monitoring, increase the proportion of time spent in the therapeutic range, and to achieve better outcomes among patients receiving warfarin therapy.[14]

## B. PATTERNS OF MISSING DATA IN THE LRT

While no formal assessment has determined patterns of missing data in the LRT, some patterns are certainly present due to the nature of clinically-derived data. For example, a *general missing data pattern* is likely common. In a general pattern, values are missing in an apparently random distribution, but values can be systematically missing. A general missing data pattern can occur for example, with provider or patient non-adherence to recommended laboratory monitoring. A *monotone missing data pattern* typically occurs with longitudinal data assessment and is seen for example, after individuals terminate insurance plan enrollment. To the extent possible, we will describe missing data patterns in MSDD examples and review statistical approaches most robust to differing missing data patterns.

## C. MECHANISMS OF MISSING DATA IN THE LRT [15]

### 1. Missing completely at random (MCAR)

MCAR data exist when the probability of missing data on a variable is unrelated to other measured variables and is unrelated to the values of the variable itself (i.e., completely unrelated to the data, random missingness). MCAR is the least problematic; however, it is unlikely in the LRT data.

### 2. Missing at random (MAR)

MAR data exist when there is no relationship between the probability of missing data on a variable and the values of that variable, after conditioning on other measured variables (i.e., the probability of missing data of a variable is solely a function of other measured variables). A possible example of MAR data in the LRT is the missing calendar year 2006 data from two DPs. In most analytic situations it is not possible to confirm that data are MAR.

### 3. Missing not at random (MNAR)

MNAR data exist when the probability of missing data on a variable is related to the values of the variable itself, even after conditioning on other measured variables. It is not possible to determine that data are MNAR without knowing the result values of the missing variables.[2]

### 4. Importance of the missing data mechanism

MCAR, MAR, and MNAR mechanisms are important because they are assumptions that govern the performance of missing data handling strategies.[2] For example, restricting to cases with complete covariate data assumes and requires MCAR data and will provide biased estimates if the data are MAR or MNAR. More robust missing data techniques, such as multiple imputation (MI), require the less stringent MAR assumption. Any missing data technique will produce biased estimates if the underlying assumptions are violated.

# IV. LITERATURE ON THE USE OF CLINICAL LABORATORY TEST RESULTS DATABASES IN CLAIMS AND ELECTRONIC HEALTH RECORD DATABASE STUDIES OF MEDICAL PRODUCT SAFETY

## A. LITERATURE REVIEW OVERVIEW

The Mini-Sentinel Analytic Methods for Using Laboratory Test Results in Active Database Surveillance Workgroup (the Workgroup) reviewed, described, summarized, and interpreted the published literature on the use of clinical laboratory test results in electronic health records database and other observational data studies of medical product safety. The main emphasis of the review was on missing laboratory test results where clinical laboratory results were included as covariates, used in cohort identification, or used in detecting health outcomes, and how missing observations were addressed in published studies, including mechanisms and patterns of missing data, predictors of missingness, analytic methods to handle missing observations, and assessment of the methods' performance. In addition, we created a database to summarize the reviewed articles and pertinent details of study design, specific laboratory tests studied, missing data descriptions, and any missing data techniques used.

Because the overall goal of this literature review was to describe missing data analytic methods used by studies comparable to those expected to be undertaken within Mini-Sentinel, the literature review informed the Workgroup by providing a critical evaluation and narrative summary of the methods used in published observational research studies using claims data, other electronic administrative data, or EHR data where missing laboratory results data were encountered. Because prospective intervention studies use data collection techniques that are not common to our focus on observational studies, we did not include them in this review. Prior to beginning the literature review, the Workgroup identified two key questions:

- Key Question 1. Regarding the actual missing clinical laboratory test results in published studies employing laboratory results databases,
    a. Were missing data acknowledged and described?
    b. Were the mechanisms of missing test results identified? If so, what were they?
    c. Were the patterns of missing test results identified? If so, what were they?
    d. Were the predictors of missingness of test results determined? If so, what were the predictors?
- Key Question 2. Regarding how missing clinical laboratory test results observations in published studies employing laboratory results databases were handled in those studies,
    a. What analytic methods were employed?
    b. Was the performance of these methods addressed (e.g., was bias assessed, were different methods compared)?
    c. If the performance of these methods was addressed, how did the methods perform?

## B. LITERATURE REVIEW METHODS

Descriptions of methods used for handling missing clinical laboratory test results in observational claims and electronic health records database studies of medical product safety are often not reflected in papers' titles, abstracts, or key words. For that reason, we did not conduct a systematic literature search

to identify studies for inclusion in the review. Instead, we conducted multiple searches using varied strategies. We also consulted investigators with expertise in missing data methods, reviewed the reference sections of papers identified as potentially relevant, and conducted an exhaustive search of papers published in one journal as part of this review. Here we describe important elements of the search process.

Literature searches were conducted by searching PubMed, OVID, and the Cumulative Index to Nursing and Allied Health Literature (CINAHL®). The National Library of Medicine's Medical Subject Headings (MeSH) keyword nomenclature developed for MEDLINE® and adapted for use in other databases was employed, with searches limited to studies published in English between January 1, 2000 and June 30, 2014. Search details, including search terms and numbers of articles identified in each search, are provided in the Table in Appendix A. Preliminary, general searches employing terms such as "missing" or "missing data" as text or key words in conjunction with laboratory indicator terms (i.e. "laboratory" as text word; "diagnostic test" or "routine" as MeSH terms) returned many thousands of citations. Unfortunately, review of the titles of these citations revealed almost none pertaining to research that analyzed laboratory results data with missing values; review of abstracts of the very few potentially pertinent citations identified most of those papers to also not be relevant (e.g., studies that gathered laboratory results via chart review).

Searches focused on specific laboratory test types and on particular missing data techniques were conducted next. In focused searches we selectively queried 11 specific laboratory test types (i.e., glucose, hemoglobin, HbA1c, platelets, alkaline phosphatase, alanine aminotransferase, bilirubin, creatinine, creatine kinase, lipase, and INR) together with a laboratory term ("laboratory" or "diagnostic test" [text words]; "routine" [MeSH term]) and either "missing" (as a text word) or selected missing data analytic technique terms (i.e.. MI, predictive mean matching, and MNAR techniques [pattern mixture, selection model], longitudinal studies). Focused searches retrieved few citations, but most citations found were pertinent.

By definition, within the context of our key questions, searches focused on a particular missing data technique had important limitations. For example, focused searches could not assist in determining relative frequencies of use of different missing data techniques or which missing data techniques were most commonly applied overall. For these reasons, we also examined the titles, abstracts, and analytic methods applied in every article published in *Pharmacoepidemiology and Drug Safety* (PDS) between 2001 and mid-2014 that examined laboratory result values, regardless of whether the article was retrieved in any of the previous searches. We chose PDS because its focus is similar to the activities undertaken within Mini-Sentinel.

We also utilized the expertise of the Workgroup by soliciting relevant articles known to Workgroup members. References of pertinent articles retrieved through all of these methods were examined to identify additional studies of potential interest.

Potentially relevant citations were imported into a relational database specifically developed for this project. A screenshot of the database entry form is shown in Appendix B. Abstracts were reviewed and the full texts of relevant articles were retrieved and further reviewed. Each study's methods and findings were summarized using a standard set of criteria (Appendix C). Examples of the criteria used to critically review and summarize each paper included the following:

- What were the study objective, population, general methods, and result?
- Were the laboratory results used in the study from one or multiple organizations?

- Does the study include results of only one laboratory test type (e.g., serum creatinine) or results from several different test types?
- What is the size of the study/number of study patients and laboratory tests (e.g., is the study population size potentially relevant to the Mini-Sentinel environment of studying rare adverse events/rare exposures)?
- How were laboratory results used (i.e., as covariates, in outcomes detection, or for cohort identification)?
- Does the study address sources of missing data?
- Are the mechanisms and patterns of missing data addressed?
- Are predictors of missingness studied?
- Are analytic methods to handle missing observations addressed? If so, what analytic method was applied? If so, were sensitivity analyses completed/ was method performance assessed?

## C. LITERATURE REVIEW RESULTS

### 1. Use of missing data techniques in publications analyzing laboratory results data

Although limited use of missing data techniques in the published literature was indirectly evident from the low capture of pertinent articles in searches, the best estimate of overall use of missing data techniques when analyzing laboratory results data resulted from the search of articles in PDS that included laboratory results data. From 2001 to mid-2014, 112 articles were identified in PDS using the laboratory search terms. The majority of these articles (77%, n=86) were not applicable to our search for observational claims and electronic health records database studies of medical product safety for the reasons noted in Table 1.

Among the 26 relevant papers from PDS, one article used HbA1c $\geq$ 7%, if available (without addressing missingness), as one pathway for cohort entry, but also used diagnosis codes or anti-diabetic medication prescriptions.[16] Three articles described missing laboratory results data but did not analyze the laboratory results data.[1, 17, 18] Eighteen of the remaining 22 articles (82% of the 22) used complete case analyses. Thus, only four articles published in PDS used missing data techniques: two MI,[19, 20] one regression imputation[21] and one inverse probability weighting.[22] These results suggest that use of missing data techniques when analyzing laboratory results data has been very limited to date; complete case analyses predominate.

**Table 1. Reasons 86 of 112 Article in Pharmacoepidemiology and Drug Safety were Not Pertinent in Database Studies of Medical Product Safety that included Laboratory Results Data**

| Reasons Articles were Not Pertinent | Number (total = 86) | % |
|---|---|---|
| Laboratory results data not evident [a] | 19 | 22 |
| Laboratory results data not analyzed [b] | 13 | 15 |
| Missing data not mentioned or described | 11 | 13 |
| Data collection technique (e.g., direct data capture, meta-analysis) | 11 | 13 |
| Article type (e.g., case study, commentary) | 10 | 12 |
| Monitoring/laboratory test completion study | 8 | 9 |
| Reference value study/Database or coding validation | 4 | 5 |
| Other [c] | 10 | 12 |

[a] Articles mentioned future study recommendations, limitations, or non-laboratory diagnostic tests
[b] Articles looked at policy impacts, decision analyses, tools for adverse event monitoring, etc.
[c] Examples: small pilots, record linkage studies, prescribing alerts

## 2. Classification of potentially pertinent articles

We found a total of 240 potentially relevant articles (112 from PDS and 128 from the other general and focused searches). As noted above, use of missing data techniques when analyzing laboratory results data was relatively rare. Nevertheless, the few articles utilizing methods to adjust for missing laboratory results data offer insights for the current project. Here we briefly describe all potentially relevant articles identified across all searches and provide detail from selected articles.

Categorization of the 240 potentially relevant articles is shown in Table 2. Upon examination of the articles, over half (n = 129; 54%) were not pertinent. Eight articles described missing data patterns but did not analyze data and had no need to apply missing data analytic techniques. However, these eight articles were useful in describing patterns of missing data (see below). Articles classified as 'Methods only' did not include laboratory data analyses. Of the 77 articles categorized as pertinent, 37 (48%) used at least one missing data technique (details on the methods used in these articles are in Section V.C.4.), 38 (49%) used complete case analyses only, and two (3%) utilized laboratory data as supplementary data for cohort selection only.

**Table 2. Classification of the 240 Potentially Relevant Articles Identified during Literature Review**

| Classification | Number (total = 240) | % |
|---|---|---|
| Not pertinent [a] | 129 | 54 |
| Used or analyzed missing laboratory results data | 77 | 32 |
| Missing data technique | 37 | |
| Complete case analysis | 38 | |
| Supplemented cohort selection | 2 | |
| Methods only | 23 | 10 |
| Described missing laboratory results data only | 8 | 3 |
| Not able to classify, full article not available | 3 | 1 |

[a] 86 of 'not pertinent' articles detailed in Table 1. For the remaining 43, the most common reasons were missing data not mentioned or described (17 [40%]), laboratory results data not analyzed (7 [16%]) and laboratory results data not evident (6 [14%])

## 3. Descriptions of uses or analysis of missing laboratory results data in published articles

Although 77 articles (Table 2) provided some evidence of missing laboratory results data, descriptions of missing laboratory results data were generally incomplete. As one example, only ten articles (14%) described missing data beyond simply noting the percentage missing. As another example, one article required laboratory results to be available for inclusion in the analyses presented but did not present a cohort selection flowchart or other description of the full population (i.e., the reader could not determine the proportion excluded due to missing laboratory results data).[23] Another published analysis[24] used last observation carried forward when analyzing outcome data at specific follow-up times, but the number and timing of follow-up data were not described. Some of the most detailed descriptions of missing laboratory results data were in articles that only described missing data and did not analyze data within the same paper. [1, 25, 26] This observation potentially suggests that limitations on article length and table number, or perhaps perceived lack of importance of missing data to the study purpose (by authors, reviewers and/or editors) contributed to lack of published details about missing data.

The following patterns and details about missing laboratory results data were noted in the publications:

- Missing data patterns described in existing publications generally focused on patient characteristics, also referred to as Patient-Level Missingness. Putting such missingness in context, patient-level missingness can be patient-driven, provider-driven, or reflect a combination of patient and provider factors. Provider decisions to order laboratory tests may strongly reflect population characteristics such as perceived risk or severity of illness, while patient decisions to adhere to completing ordered tests can reflect characteristics such as low health literacy, access issues, or competing priorities, for example. Most published observational database studies did not – could not-- differentiate between whether a) the data were missing because the provider did not order the test or b) the patient did not adhere to completing the ordered test. This is because currently published studies employing clinical laboratory results did not have access to test *orders*, but only could access *completed* tests. Additionally, if a laboratory result is not necessary to the clinical care of the patient, but a researcher expects the laboratory result to be available for research purposes, the test result is considered to be "missing." However, this "missingness" when a test is not done is not missing in the same sense that a test result is missing when the test was completed and, for some reason, that test result is not available within the database.
- Examples of Patient-Level Missingness noted in publications:
    - Two papers reported that missing laboratory results values occurred more often among persons who were younger and had fewer comorbidities,[27, 28] while two other publications reported persons excluded due to missing data were older with a higher probability of mortality in the following years[29] or had mixed comorbidity patterns.[30]
    - Witt and colleagues[31] noted that missed INR tests results, i.e. non-adherence with ordered INR testing, likely also reflected non-adherence with prescribed warfarin use.
- When looking at dyslipidemia in youth, Li et al.[32] found children with lipid testing were much more likely to be obese, have diabetes, or to have hypertension likely reflecting selective ordering by providers. Such patterns may be accentuated when frequency of testing in general is low, as in their study where only 7% of children had any lipid laboratory results available.
    - Pivovarov and colleagues examined temporal measurement patterns, specifically identifying short gaps versus long gaps between tests as one tool for indicating different test patterns related to distinct disease states.[25] Varied patterns emerged for particular laboratory tests, but as one example, high lipase values were more strongly associated with an acute pancreatitis diagnosis when gaps between laboratory measures were short (0-3 days) compared to values with gaps > 3 days. This example reminds us that laboratory tests are ordered, completed, and interpreted in specific contexts (e.g., diagnosis of acute illness, monitoring of response to therapy or monitoring disease progression) during episodes of providing healthcare and that the care context is often not obvious in observational database studies.
- Very few publications discussed missing data patterns related to system, location, or temporal factors, that is Facility-Level Missingness, Organization-Level Missingness, Care-Setting Level Missingness, and/or Temporal-Level Missingness. However, these are important types of "missing" data for Mini-Sentinel activities and are common missing data types in the MSCDM

because these missing data types reflect situations where the test was completed but, for some reason, the test result is not available within the database.

- Trickey and colleagues[33] noted data availability differed between trauma centers due to healthcare site processes and policies. Similarly, Raebel and colleagues [1] reported rates of metabolic monitoring of youth starting second-generation antipsychotic agents differed between integrated and less-integrated healthcare systems in the United States; these differences remained significant after accounting for selected patient characteristics.[34] Health plan type may also impact healthcare use and data capture. For example, Reddy et al[35] reported a modest reduction in general laboratory testing for persons who were switched to High Deductible Health Plans compared to persons who remained in traditional HMO plans.
- One of the most extensive descriptions of laboratory data from administrative sources examined available low-density lipoprotein (LDL) cholesterol, high density lipoprotein (HDL) cholesterol, and HbA1c tests in a large cohort of patients initiating lipid lowering therapy at 14 Blue Cross/Blue Shield health plan sites in the US.[26] This work by Schneeweiss and colleagues captured outpatient laboratory results and claims codes (a claim code indicates test completion but does not provide the test result value). In their cohort, 68% of patients had a laboratory test claim, while 42% had laboratory results available. Whether a patient had evidence of a laboratory test being completed was strongly associated with patient characteristics, while laboratory results availability varied by healthcare system characteristics and state of residence. They highlighted several additional factors that impacted laboratory data capture including: a) lower testing rates for persons recently hospitalized who may have had tests completed as inpatients; b) patients with Medicare Supplemental coverage had a lower proportion of claims and few results available attributed to health plans typically being secondary, not primary, payers; c) laboratory results for their cohort were only available from specific contracted laboratory providers who did not operate in all the states studied; and d) patients receiving lipid treatment after recent myocardial infarction or acute coronary events had lower rates of laboratory testing, probably due to receiving the drug as secondary prevention. Of note, this study was conducted using the databases of a Mini-Sentinel DP that is a large national insurer.

In summary, the findings of patterns of missing laboratory data in published studies highlight three main points. First, consistent patterns of missing data should not be expected because missing data patterns are heavily influenced by the specific laboratory test type(s) of interest, by study population characteristics, and by provider and system factors. Therefore, describing missing data overall and by key variables within the study population is an important early step in research studies. Second, the existing literature is of little assistance in informing the activities of this Workgroup about handling Facility-Level, Organization-Level, Care-Setting Level, and Temporal-Level Missingness. Third, the complexities of reasons for missing laboratory results data deserve thoughtful discussion during analysis planning. The assumption of MCAR, a requirement for unbiased complete case analyses, rarely holds for

[1] This paper is not included in the number of articles identified during the search period through June 2014 (i.e., is not included in information in the Tables) because it was published after June 2014. It was available to the Workgroup as a draft manuscript at the time the literature search was conducted.

missing laboratory results data. MAR patterns may be found but are expected to be difficult to confirm with missing laboratory results data. Whether there is evidence of a strong MNAR pattern should be considered.

## 4. Missing data techniques used for laboratory results data

As mentioned previously (Table 2), we identified 37 articles that used at least one missing data technique. The technique applied in the analyses used in each of these papers is shown in Table 3.

**Table 3. Types of Missing Data Techniques used for Missing Laboratory Test Results in 37 Studies**

| Missing Data Technique (Number of Studies) | Reference(s) | Application of Technique [b] | | | |
|---|---|---|---|---|---|
| | | Independent variable(s) | | | Dependent variable |
| | | Confounder | Exposure | Predictive Model | Outcome Identification |
| Multiple imputation (MI) [c] | | | | | |
| MI by joint modeling (10) | 19, 20, 30, 33, 36-41 | 6 | | 4 | |
| MI by Fully Conditional Specification (7) | 27, 42-47 | | 2 | 4 | 1 |
| MI with predictive mean matching (3) | 48-50 | 2 | | | 1 |
| Unspecified MI type (7) | 29, 51-56 | | 2 | 4 | 1 |
| Missing category (4) | 39, 56-58 | 2 | 2 | | |
| Mean/median substitution for missing (3) | 27, 43, 57 | 1 | 1 | 1 | |
| Last observation carried forward (3) | 24, 59, 60 | | | | 3 |
| Inverse probability weighting (2) | 45, 61 | | | | 2 |
| Propensity score calibration (1) | 30 | 1 | | | |
| Regression (single value predicted) (1) | 21 | | | | 1 |
| Hot Deck imputation (1) | 58 | 1 | | | |
| Pattern mixture model (1) | 60 | | | | 1 |
| Other (6) | 30, 33, 42, 43, 62, 63 | 1 | 1 | 3 | 1 |

[a] Do not total to 37 because some studies applied more than one technique
[b] Only 2 articles used laboratory results for cohort identification; neither used a missing data method
[c] Limited statistical information on details of MI in some articles; specific types of imputation methods noted only if specifically stated by authors, remainder categorized as Unspecified MI type

More detailed information about all 37 studies, such as the study designs, populations, laboratory test result(s) studied, % missing, missing data method used, and whether the authors described the missingness mechanism, is provided in the project-specific relational database (Appendix C).

### a. Selected examples of applications of missing data techniques

Twenty-two of the 37 articles analyzed the data using more than one method. Most often this occurred when a missing data analysis method was used in addition to a complete case analysis (12 of 22; 55%). Several articles used the complete case analyses for the primary result while employing a missing data technique in secondary analyses to confirm results remained consistent.[29, 40, 42, 48] Raebel et al[42] examined predictors of antihyperglycemic drug initiation using complete case analyses but also completed MI analyses to confirm results remained unchanged if persons with missing smoking, race, or serum creatinine test results at baseline were included. To further confirm missing covariates were not overly influential, they also completed sensitivity analyses that assigned values for missing covariates to the category with the largest effect, or the least effect, as well as the most common referent category while observing impacts on the relative risk estimates. In contrast, other authors reported primary analyses using a missing data technique while using complete case analysis for sensitivity analyses.[38, 41,

[59] For example, Nakano et al[41] presented primary analyses with MI and used complete case analyses for sensitivity checks.

A few authors more fully presented results from two or more missing data methods. For example, Harris et al[45] presented adjusted results for models using both inverse probability treatment weights and MI to retain the 54% of their population who were missing HbA1c result values. They believed fully reporting the similar results obtained from both methods supported the validity of their inferences. However, most authors that fully presented results using two or more different missing data methods did so to teach or encourage use of the method. Examples include:

- Mulla et al[37] sought to encourage use of MI through demonstration of the technique. They presented models predicting an outcome of hospital mortality in a population where serum albumin was available only for 55% of the cohort. Models were run for complete case analyses (N=110) and a full dataset (N=201) with serum albumin imputed using MI. Higher serum albumin was similarly protective in both the complete case and MI models; however age 55 years or older (compared to age 0 – 54 years) was not a risk factor for hospital mortality in the complete case analysis (adjusted odds ratio [OR] 2.43, 95% confidence interval [CI] 0.79-7.53), but was a significant risk factor in the imputed cohort (OR 3.08, 95% CI 1.22 – 7.78).
- Faries and colleagues[30] examined three analytic methods to adjust for HbA1c when it was only available for 25% of their cohort. Their analyses examined Bayesian modeling, propensity score (PS) calibration, and MI. Adjustments for baseline HbA1c using Bayesian modeling and MI provided similar results and neither was far from the result of the original analyses, suggesting the added adjustment had a limited impact. They noted that the assumptions for PS calibration were questionable in their analyses because the subsample with HbA1c was not random and the estimated PS were not independent of the outcome after accounting for the subsample PS and treatment. Nevertheless, results remained reasonably similar to the other two methods.

**b. Precision versus accuracy in missing data techniques**

Comparing different missing data techniques can provide some information as to whether answers are similar or different (precision), but it can be difficult to ascertain which answer is closest to the truth (accuracy). To circumvent this issue, some studies have created missing data.

- Siew and colleagues[27] examined different methods of accounting for missing baseline serum creatinine results values in models predicting Acute Kidney Injury (AKI). They first created models predicting which patients had missing data and used those models to identify similar patients among persons with serum creatinine results values available. Removing creatinine values for those patients gave them MAR data which they analyzed using both a single imputation and an MI method. The availability of actual laboratory results values allowed them to compare mean differences between the imputed and actual results as well as to contrast differences in the prediction of AKI (although artificially creating missingness does not necessarily reflect real-world patterns of missing data). They concluded MI generally improved accuracy of predicting AKI.
- Using somewhat similar methods, Walijee and colleagues[43] randomly removed laboratory results data from a dataset with varying frequencies and compared accuracy of imputations using simple mean imputation, imputation by nearest neighbor, MI by Fully Conditional Specification (FCS; also known as chained equations) and imputation using random forest models. They reported random forest methods had the lowest imputation error and the smallest prediction differences

when imputed laboratory values were used. MI with FCS had the second lowest imputation error and prediction differences, while larger discrepancies were seen for nearest neighbor and mean imputation. They acknowledged that limitations to the random forest models were a requirement for skilled R programming for implementation and higher computational needs compared to the MI by FCS.

### c. Multiple imputation approaches

MI is a relatively common missing data technique that starts with the observed incomplete data and then creates multiple complete versions of the data by replacing missing data with varying plausible values. Several different imputation methods exist, e.g., MI by joint modeling, by fully conditional specification, and with predictive mean matching. Predictive mean matching is one MI method that ensures the replacements reflect observed data values. We located three articles using predictive mean matching.[48-50] Morris and colleagues[49] contrasted predictive mean matching to local residual draws in extensive simulations as well as in an example with missing albumin data in models predicting time to death among persons with ovarian cancer. In their data example, the model results were comparable for replacements using predictive mean matching versus the related method of local residual draws. Their simulations also showed these two imputations methods performed better than parametric imputation when the imputation model was mis-specified but results were still poor for mis-specified models for all methods, particularly when missing covariates were strongly associated with the outcome.

### d. Missing not at random concerns

Although articles describing missing laboratory results data suggest some patterns that may not be missing at random (MNAR), we found no articles employing missing data techniques better able to account for MNAR patterns when analyzing missing laboratory results data. The literature search identified one methods article that employed pattern mixture models to examine missed visits that resulted in interval-censoring of total bilirubin outcomes.[60] While it is possible we could have missed applied articles using MNAR techniques in our search, it seems appropriate to conclude that there likely has been very limited use of such techniques.

## V. CONSIDERATIONS OF A TWO PHASE DESIGN APPROACH WITH MISSING LABORATORY DATA IN MINI-SENTINEL

While not a part of this literature review activity, the report from a previous Mini-Sentinel Methods Workgroup, the Two-Phase Sampling Workgroup (www.sentinelinitiative.org/sentinel/methods/327) was reviewed and discussed to determine relevance to this missing laboratory results data methods project. This previous Workgroup considered the use of Two-Phase designs for Mini-Sentinel medication safety surveillance activities.

The work of a two-phase design entails using phase 1 data, the information available on everyone, to identify a phase two targeted subgroup for collection of supplemental confounder information from alternative sources such as medical records. That Workgroup's report primarily focused on the collection of supplemental confounder data using a Two-Phase design, but also discussed using such a design for exposure and outcome ascertainment. A necessary step in the development of a Two-Phase design is to create an investigator-designed, probability-sampled sub-cohort that can be used to collect additional confounder information (i.e., in phase 2) via supplemental means such as chart review. Phase 2 findings

are then translated to the entire cohort on the basis of the investigator-designed probability sampling. The investigator can substantially reduce the risk of selection bias in the sub-cohort using such methods, in part because the sampling probabilities are known to them.

The Two-Phase Workgroup's authors make a careful distinction between this deliberate sampling scenario used with a Two-Phase design and what they termed 'opportunistic' supplemental data collection wherein investigators take advantage of electronic data available at some, but not all, DPs. As those authors note, laboratory data fits this definition of opportunistic data. Unlike the probabilistic, deliberate nature of the sampling and data collection in a Two-phase design, a cohort with opportunistic data have those data available for reasons that may not be known by the investigator, thus there is an increased risk of selection bias. Opportunistic data may exist for reasons unknown to, and beyond the control of, the investigator. For example, opportunistic data may be linked to care delivery methods (e.g. fee for service vs. managed care), patient mix, and structural and financial issues such as payment and carve-outs, etc. Thus, much of the findings and recommendations from the Two-phase report are not directly transferrable to Laboratory Methods Workgroup when using opportunistic data. However, the completeness of potential confounders available in phase one data regarding the information available on everyone is critical in providing an unbiased approach to identify the phase two targeted subgroup. To the degree that a study design utilizes laboratory data either coded procedures and/or availability of a laboratory test result to subset the population for phase two, the imputation methods employed by the Laboratory Analytic Methods Workgroup may inform the sampling strategies employed in Two-phase studies.

There are scenarios under which the findings from the Two-Phase Workgroup are directly transferable to the Laboratory Analytic Methods Workgroup:

- When laboratory results are not available electronically from any Data Partners; this might be a specialty test that is done only at facilities with particular equipment, perhaps some forms of genetic testing. In those cases a Two-Phase design could be used, with the laboratory data being obtained in the second phase using primary data collection.
- The design of a two-phase study may enhance the ability to confirm findings in employing imputation methods. While one would likely not need to perform a validation study confirming that an electronically recorded laboratory test result matches a value in a patient's chart, the characterization of the mechanism of missingness of a laboratory result (e.g., test not performed versus test performed but recorded into a faxed document) could be aided by chart review of a subset of the population with missing measurements. Observational study designs make certain assumptions such as new users of an oral antidiabetic agent will receive frequent healthcare and regular laboratory monitoring. However, electronically missing laboratory procedures or results may not be an indicator of poor care but reflect the underlying nature of medical documentation in the setting or the extent of access to the source laboratory results data. The characterization and imputation methods employed by the Laboratory Analytic Methods Workgroup could be followed up with a two-phase approach to sample charts of imputed patients to confirm strategies.

Characterizing the population as part of the imputation methods employed in the activity of the Laboratory Analytic Methods Workgroup helps inform the selection bias noted in "opportunistic" sampling noted in the two-phase report. Thus the characterization of missing laboratory data may provide clarity to the question of why data are available for certain people and not others. There are methodological challenges in utilizing already collected data in opportunistic sampling as which and how

many people to sample is not chosen by the investigator but by the data available and the distribution of data may or may not target the informative people.

# VI. APPLICATION OF MISSING LABORATORY DATA EXAMPLES FROM PUBLISHED LITERATURE TO MINI-SENTINEL

This literature review focused on published articles most likely to inform future Mini-Sentinel analyses. We did not include articles where missing clinical results data would be prospectively minimized (e.g., studies with direct data capture or closely monitored clinical trials). In this literature review of observational database studies, few studies included very large populations or included multiple sites. As a result, applying information from the articles identified in the literature review to Mini-Sentinel activities must be done cautiously. However, given this caution, MI methods emerged as important candidates for Mini-Sentinel work, particularly given the depth of prior methodological work and increasing availability in software packages. From our review, it is also clear that methods with known limitations, such as simple mean or median replacement of missing values that result in inappropriately small standard errors, should not be considered for Mini-Sentinel activities.

# VII. EVALUATION OF ANALYTIC METHODS FOR INCORPORATING LABORATORY TEST RESULTS DATA IN MINI-SENTINEL MEDICAL PRODUCT SAFETY ANALYSES

## A. GENERAL PROCESS FOR TESTING THE FUNCTIONING AND COMPUTATIONAL PERFORMANCE OF THE MISSING DATA ANALYTIC APPROACHES

Missing data approaches have been extensively detailed for decades in books and articles and new methods continue to be developed. For this project that included assessing the performance of missing data analytic approaches primarily within the context of test cases that employ laboratory test results data we selected methods that could be readily implemented using available software. For the Baseline Confounder Test Cases, more than one missing data method was implemented for each test case, with some methods repeated across test cases to aid interpretation and understanding of the impact of missing data approaches on the analysis. For all test cases, we decided on missing data techniques after viewing descriptions of missing data in the test case cohort and considering options. We first investigated predictors of missing laboratory test results values overall and by site. In test case outcome models, we typically compared results across a) analyses that did not include laboratory results at all, b) analyses that only included available laboratory results (i.e. complete case analyses), and c) analyses that accounted for missing laboratory results using varying techniques. Our primary comparisons were the estimated coefficients and 95%CI for the exposures of interest. One question that we answered for each test case was whether including laboratory results made any difference. If the missing data analytic method did make a difference, we determined the ways in which the results varied by method. In addition, the ease of implementation and the plausibility of assumptions were taken into account when making recommendations.

## B. SELECTION OF SPECIFIC TEST CASES

Given that the sources of laboratory test result missingness could vary across different laboratory test types, a predefined goal in selecting test cases was to develop generalizable information (i.e., information not just useful for one laboratory test). Test cases were also selected based on topics identified by the FDA as of interest and relevance to Mini-Sentinel activities (e.g., PROMPT, Clinical Data Elements, and other Workgroups); test cases were chosen in consultation with the FDA investigator members of the Workgroup. Further, the test cases were selected based on laboratory test types that are available in the LRT.

Importantly, the Workgroup selected test cases expected to represent different missing data issues. For example, an outcome test case was chosen to enable assessment of the absence of inpatient data for an outcome usually associated with an emergency department visit or hospitalization. A baseline confounder adjustment test case was chosen to enable assessment of incomplete outpatient data capture such as for an ambulatory laboratory test that was anticipated to have differential results missingness across DPs. Another test case was chosen to represent LRT data expected to have the minimum possible missing data issues. Such a scenario was important to include as it potentially represents the "best case." In Table 4 we list the medical product-outcome pair test cases selected to assess in this activity.

**Table 4. Selected Medical Product-Outcome Pairs Test Cases**

| Laboratory Test Result | Medical Product Exposure | Outcome | Expected Main Missing Data Issue(s)/ Expected Availability in Mini-Sentinel Distributed Database |
|---|---|---|---|
| **Baseline Confounder Adjustment** | | | |
| Baseline Confounder Test Case 1: Baseline HbA1c or fasting blood glucose or random blood glucose (composite variable "GLU") | Second generation antipsychotic (SGA) newly-started in adults who do not have a diabetes diagnosis | Diabetes diagnosis | Patient or provider non-adherence to recommended laboratory monitoring; organization-level differential missingness by data partner type; potentially some of laboratory test types with least missing data because often routinely obtained as part of usual care |
| Baseline Confounder Test Case 2: Baseline serum creatinine in patients with diabetes starting an angiotensin converting enzyme inhibitor (ACEi) | Angiotensin converting enzyme inhibitor (ACEi) initiation in patients with existing diabetes diagnosis | Hyperkalemia diagnosis [a] | Organization-level differential missingness by data partner type; patient or provider non-adherence to recommended laboratory monitoring; potentially one of the laboratory test types with the least missing data because often routinely obtained as part of usual care |
| Baseline Confounder Test Case 3: Baseline INR in current warfarin users starting an antimicrobial medication | Antibiotic initiation in patients taking warfarin chronically | Bleeding diagnosis | Organization-level differential missingness by data partner type; potentially one of the laboratory test types with the least missing data because routinely indicated in cohort patients taking warfarin |
| **Cohort Identification** | | | |
| Cohort Identification Test Case 1: Enhanced identification of a pregnancy cohort: Use of positive qualitative or high quantitative pregnancy test results | Not required | Not required | Patient-level missingness (e.g., selective assessment correlated with high risk pregnancy) |

| Laboratory Test Result | Medical Product Exposure | Outcome | Expected Main Missing Data Issue(s)/ Expected Availability in Mini-Sentinel Distributed Database |
|---|---|---|---|
| Cohort Identification Test Case 2: Enhanced identification of a cohort of adults with Chronic Kidney Disease (CKD): Use of serum creatinine laboratory results data to estimate patients' glomerular filtration rate | Not required | Not required | Organization-level differential missingness by data partner type; patient or provider non-adherence to recommended laboratory monitoring; potentially one of the laboratory test types with the least missing data because often routinely obtained as part of usual care |
| **Outcome Detection** | | | |
| Outcomes Detection Test Case 1: Diabetes and blood glucose or glycosylated hemoglobin laboratory test results values outcomes among adults initiating an SGA | SGA newly-started in adults | Hyperglycemia or diabetes diagnosis and/or elevated HbA1c, random glucose, or fasting glucose test results | Patient or provider non-adherence to recommended laboratory monitoring; organization-level differential missingness by data partner type; potentially some of the laboratory test types with the least missing data because often routinely obtained as part of usual care |
| Outcomes Detection Test Case 2: Upper gastrointestinal bleeding and hemoglobin (Hgb) laboratory test results value outcomes among adults initiating a non-steroidal anti-inflammatory drug (NSAID) | NSAIDS | Acute gastrointestinal bleeding diagnosis and/or low Hgb test results | Care setting level missingness (patient location where laboratory testing was conducted and absence of hospital and emergency department laboratory test results) |
| [a] Serum potassium result values not in Mini-Sentinel Distributed Database | | | |

## C. DATA PARTNER INVOLVEMENT AND DATA DEVELOPMENT APPROACH

Three representative Data Partners provided data for test cases. These Data Partners included one smaller integrated delivery system ("site 1"), one larger integrated delivery system ("site 2"), and one large national insurer ("site 3"). Sites 1 and 2 provide care delivery and insurance for their members, employ an electronic health record in all ambulatory medical offices where laboratory test orders and results are recorded as part of the care process, have laboratory facilities in each medical office, and have integrated laboratory results databases. Sites 1 and 2 have administrative claims data, and have access to outpatient and inpatient laboratory test results. Site 2 also has access to emergency department laboratory test results. Site 3 provides insurance only, augmenting its administrative claims data by contracting with some national laboratory service vendors to obtain clinical laboratory test results for its enrollees. At the time of this project, the sites' enrollments ranged from about 500,000 to many million. Site 3 has a larger proportion of older enrollees than do Sites 1 or 2.

Each test case essentially was a separate cohort study. The Mini-Sentinel Cohort Identification and Descriptive Analysis (CIDA) tool was employed for test case development whenever feasible. Because some test cases employed the CIDA tool and others could not, the sequence of applying inclusion and exclusion criteria across the cohorts in the separate test cases differed. Additional distributed programming code to extract data for the test cases was written at Kaiser Permanente Colorado (KPCO), the lead site for this activity. The programming code written at KPCO was tested and quality-checked in

accordance with Mini-Sentinel policies and procedures. All distributed code was sent to the participating Data Partners according to Mini-Sentinel policies and procedures. Data partner sites executed the work plans against their MSDD data and returned the resulting site-specific test case datasets to the Mini-Sentinel Operations Center and KPCO where the datasets from the participating sites were combined to yield one analytic dataset for each test case. The combined datasets were quality checked and the data described.

## D. BASELINE CONFOUNDER TEST CASES

The purpose of the Baseline Confounder Test Cases was to identify analytic strategies that enable inclusion of laboratory test result data into medical product safety analyses for improved confounding control within the MSDD. This included identifying missing data analytic issues across laboratory test types as well as availability and nature of missingness of laboratory results and whether and how analytic strategies might need to differ by type of laboratory test result. Strategies such as selected sensitivity analyses were also considered.

### 1. Baseline Confounder Test Case 1: Baseline HbA1c/glucose in new adult users of a second generation antipsychotic and risk of diabetes

#### a. Baseline Confounder Test Case 1 cohort development

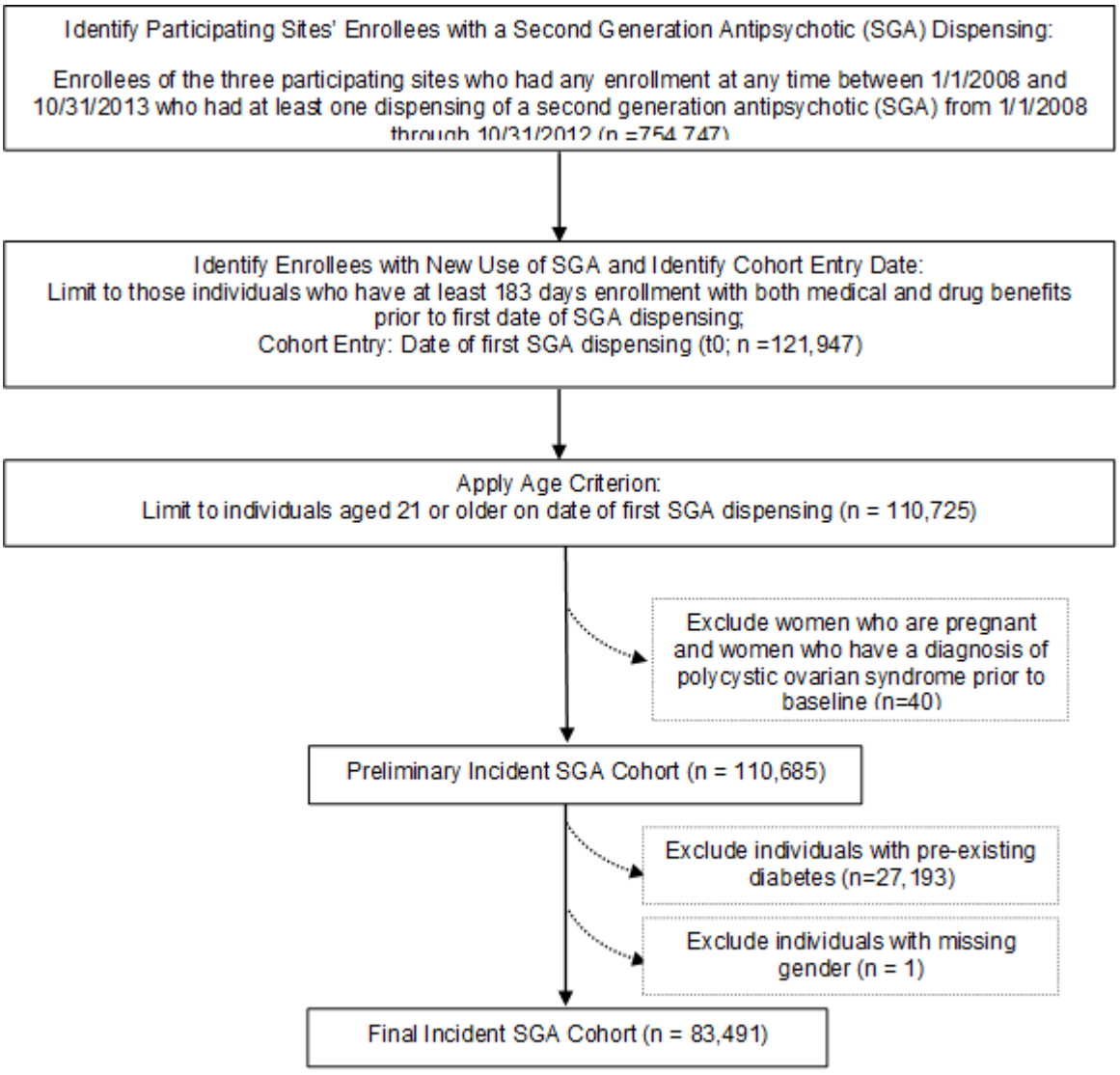
Second generation antipsychotic (SGA) agents are prescribed to aid in treatment of schizophrenia and other mental health disorders. Known risks with SGA include weight gain and the potential for developing metabolic disorders such as diabetes mellitus.[64] These risks are considered to differ across individual SGA.[65] Because studies have found worsening glucose control and new-onset diabetes during treatment with SGAs[66-68] obtaining a glucose laboratory test result at initiation of therapy is important to establish whether glucose control is normal or impaired at baseline.

The underlying premise of Test Case 1 was that FDA might wish to compare the risk of new-onset diabetes among users of different SGAs, with baseline glucose value as a potential confounder. This test case examined models that included adjustment for baseline blood glucose result value in analyses comparing initiation of several SGAs and an outcome of diagnosed diabetes. As with all test cases in this report, blood glucose result values were not consistently available; the primary goal was to describe varied methods of dealing with the missing laboratory test results data. Additional detail on the research questions and the cohort are provided below.

- Research Questions
    - Does inclusion of a baseline GLU (random glucose, fasting glucose, or HbA1c) laboratory test result value (imputed when missing) reduce the potential bias when modeling the associations between specific SGA exposure and risk of diabetes outcome in the MSDD?
    - Does inference about the relationship between specific SGA exposure and risk of diabetes outcome differ when controlling for the GLU laboratory test result under two different imputation methods (regression and predictive mean matching)?
    - Does inference differ if the data from all 3 sites are pooled and analyzed (both for imputation and the outcome model) versus if the analysis is done at each site separately, with the results combined using meta-analysis?

As detailed in **Figure 1**, this test case included persons aged > 21 years who initiated an SGA between July 1, 2008 and October 31, 2012. We identified new users by requiring a minimum of six months enrollment prior to the first dispensing of an SGA. To examine the outcome of a new diagnosis of diabetes within one year, we selected persons without a history of pre-existing diabetes and additionally excluded women with conditions that impact glucose levels (pregnancy, polycystic ovarian syndrome).

**Figure 1. Test Case Cohort for Baseline Confounder Adjustment Test Case 1, Initiation of Second Generation Antipsychotics and Missingness of Blood Glucose/HbA1c Laboratory Test Results***




*Definitions and codes are in Appendix D

### b. Baseline Confounder Test Case 1 descriptive analysis of missing laboratory results

Detailed descriptions of the cohort by SGA and by site are in Tables 5 and 6. Overall, the most common SGA used was quetiapine (48%), followed by risperidone (24%), with lowest use for olanzapine (11%). Drug use varied by site. For example, site 2 had the strongest predominance of quetiapine use while quetiapine and risperidone were used with comparable frequency at site 1. The cohort averaged 60 years of age and 62% were female. During the year following SGA initiation, 5.3% developed diabetes (as defined using diabetes diagnosis codes and anti-diabetic medication dispensing); rates were similar by type of SGA (range: 4.6% to 6.1%). Forty-one percent of the cohort had at least one baseline laboratory test result for glucose or HbA1c with moderate differences by SGA type (33.9% to 43.5%).

Availability of laboratory tests was disparate by site (**Table 6**). Although Current Procedural Terminology (CPT) codes suggested similar rates of testing across sites, test results were only available for 27.7% of site 3 cohort members versus 58.0-58.9% for sites 1 and 2. In addition, at site 3 nearly all glucose results were classified as random because of unknown fasting status. Because of this, we combined fasting and random glucose values at all three sites. We additionally confirmed that ~90% of those with HbA1c also had a baseline glucose result available and we used only glucose results in analyses.

**Table 5. Characteristics of Individuals in the Baseline Confounder Test Case 1 Population: Initiation of Second Generation Antipsychotics and Missingness of Blood Glucose Laboratory Test Results**

| Characteristics | Second Generation Antipsychotic (SGA) | | | | |
|---|---|---|---|---|---|
| | Aripiprazole N=14,588 (17.5%) | Olanzapine N=9094 (10.9%) | Quetiapine N=40,200 (48.2%) | Risperidone N=19,609 (23.5%) | Total N=83,491 |
| Female sex | 10,204 (69.9) | 5284 (58.1) | 24,408 (60.7) | 11,993 (61.2) | 51,889 (62.1) |
| Age in years, mean (SD) | 50.1 (15.6) | 61.1 (20.2) | 62.1 (20.5) | 64.2 (20.9) | 60.4 (20.4) |
| Year of cohort entry | | | | | |
| 2008 | 1818 (12.5) | 1374 (15.1) | 5210 (13.0) | 2448 (12.5) | 10,850 (13.0) |
| 2009 | 3877 (26.6) | 2106 (23.2) | 9138 (22.7) | 4446 (22.7) | 19,567 (23.4) |
| 2010 | 3425 (23.5) | 1878 (20.7) | 8902 (22.1) | 4423 (22.6) | 18,628 (22.3) |
| 2011 | 2838 (19.5) | 1900 (20.9) | 8939 (22.2) | 4397 (22.4) | 18,074 (21.7) |
| 2012 (1/1 – 10/31) | 2630 (18.0) | 1836 (20.2) | 8011 (19.9) | 3895 (19.9) | 16,372 (19.6) |
| Site | | | | | |
| 1 | 459 (3.2) | 228 (2.5) | 1564 (3.9) | 1600 (8.2) | 3851 (4.6) |
| 2 | 5320 (36.5) | 3557 (39.1) | 16,986 (42.3) | 5756 (29.4) | 31,619 (37.8) |
| 3 | 8809 (60.4) | 5309 (58.4) | 21,650 (53.9) | 12,253 (62.5) | 48,021 (57.5) |
| Outcome: diabetes diagnosis after SGA initiation [a] | 666 (4.6) | 479 (5.3) | 2091 (5.2) | 1197 (6.1) | 4433 (5.3) |
| Baseline confounder: any blood glucose result (fasting or random glucose, HbA1c) | 4949 (33.9) | 3970 (43.7) | 17,484 (43.5) | 8022 (40.9) | 34,425 (41.2) |
| Fasting glucose | 1505 (10.3) | 892 (9.8) | 3973 (9.9) | 1631 (8.3) | 8001 (9.6) |
| Random glucose | 3712 (25.4) | 3492 (38.4) | 15,020 (37.4) | 6945 (35.4) | 29,169 (34.9) |
| HbA1c [b] | 767 (5.3) | 506 (5.6) | 2178 (5.4) | 1065 (5.4) | 4516 (5.4) |
| Baseline procedure code (Glucose CPT code) | 4344 (29.8) | 4595 (50.5) | 17,521 (43.6) | 8365 (42.7) | 34,825 (41.7) |
| Hispanic ethnicity (Y vs N or unknown) | 728 (5.0) | 524 (5.8) | 2378 (5.9) | 1165 (5.9) | 4795 (5.7) |
| Race | | | | | |
| White | 9258 (63.5) | 6376 (70.1) | 29,702 (73.9) | 14,355 (73.2) | 59,691 (71.5) |
| Black | 702 (4.8) | 680 (7.5) | 2936 (7.3) | 1949 (9.9) | 6267 (7.5) |
| Other | 469 (3.2) | 477 (5.2) | 1413 (3.5) | 790 (4.0) | 3149 (3.8) |

| Characteristics | Second Generation Antipsychotic (SGA) | | | | |
|---|---|---|---|---|---|
| | Aripiprazole N=14,588 (17.5%) | Olanzapine N=9094 (10.9%) | Quetiapine N=40,200 (48.2%) | Risperidone N=19,609 (23.5%) | Total N=83,491 |
| Unknown | 4159 (28.5) | 1561 (17.2) | 6149 (15.3) | 2515 (12.8) | 14,384 (17.2) |
| Number of unique medication classes dispensed during baseline, mean (SD) | 10.8 (4.8) | 9.8 (5.3) | 10.1 (4.9) | 9.5 (4.8) | 10.1 (4.9) |
| Number of ambulatory medical visits during baseline, mean (SD) | 13.0 (15.3) | 15.0 (18.4) | 14.7 (19.2) | 13.6 (16.4) | 14.2 (17.9) |
| Emergency department visit during baseline, N (%) yes | 3785 (25.9) | 4192 (46.1) | 16,190 (40.3) | 7698 (39.3) | 31,865 (38.2) |
| Hospitalization during baseline, N (%) yes | 1873 (12.8) | 3322 (36.5) | 11,287 (28.1) | 5698 (29.1) | 22,180 (26.6) |
| Institutional stay during baseline, N (%) yes | 2030 (13.9) | 2554 (28.1) | 9271 (23.1) | 5090 (26.0) | 18945 (22.7) |
| Comorbidity score, mean (SD) [c] | 1.2 (1.5) | 2.3 (2.6) | 2.0 (2.4) | 2.2 (2.4) | 1.9 (2.3) |
| Gagne et al individual comorbidities [c] | | | | | |
| Alcohol abuse | 905 (6.2) | 773 (8.5) | 3649 (9.1) | 1340 (6.8) | 6667 (8.0) |
| Anemia, deficiency | 1245 (8.5) | 1619 (17.8) | 6314 (15.7) | 3274 (16.7) | 12,452 (14.9) |
| Arrhythmia | 812 (5.6) | 1471 (16.2) | 6077 (15.1) | 3104 (15.8) | 11,464 (13.7) |
| Coagulation disorder | 170 (1.2) | 338 (3.7) | 1251 (3.1) | 587 (3.0) | 2346 (2.8) |
| Congestive heart failure | 533 (3.7) | 1119 (12.3) | 4501 (11.2) | 2430 (12.4) | 8583 (10.3) |
| Diabetes, complicated [d] | < 6 (0.0) | 8 (0.1) | 33 (0.1) | 16 (0.1) | 62 (0.1) |
| Dementia | 446 (3.1) | 1629 (17.9) | 7893 (19.6) | 5106 (26.0) | 15,074 (18.1) |
| Fluid/Electrolyte disorders | 961 (6.6) | 1877 (20.6) | 6578 (16.4) | 3286 (16.8) | 12,702 (15.2) |
| HIV/AIDS | 69 (0.5) | 36 (0.4) | 178 (0.4) | 51 (0.3) | 334 (0.4) |
| Hypertension | 4428 (30.4) | 4274 (47.0) | 19,070 (47.4) | 9900 (50.5) | 37,672 (45.1) |
| Hemiplegia | 126 (0.9) | 163 (1.8) | 738 (1.8) | 333 (1.7) | 1360 (1.6) |
| Liver disease | 335 (2.3) | 282 (3.1) | 1133 (2.8) | 450 (2.3) | 2200 (2.6) |
| Metastatic cancer | 79 (0.5) | 308 (3.4) | 663 (1.6) | 238 (1.2) | 1288 (1.5) |
| Psychosis | 10,083 (69.1) | 5345 (58.8) | 19,807 (49.3) | 11,394 (58.1) | 46,629 (55.8) |
| Pulmonary disease, chronic | 2178 (14.9) | 1910 (21.0) | 7579 (18.9) | 3877 (19.8) | 15,544 (18.6) |
| Pulmonary circulation disorder | 93 (0.6) | 183 (2.0) | 804 (2.0) | 419 (2.1) | 1499 (1.8) |
| Peripheral vascular disease | 549 (3.8) | 953 (10.5) | 4023 (10.0) | 2255 (11.5) | 7780 (9.3) |
| Renal failure | 458 (3.1) | 885 (9.7) | 3961 (9.9) | 2336 (11.9) | 7640 (9.2) |
| Tumor, any | 562 (3.9) | 863 (9.5) | 2909 (7.2) | 1391 (7.1) | 5725 (6.9) |
| Weight loss | 140 (1.0) | 491 (5.4) | 1374 (3.4) | 610 (3.1) | 2615 (3.1) |
| Additional comorbidities specific to baseline confounder test case 1 | | | | | |
| Myocardial infarction | 281 (1.9) | 436 (4.8) | 2031 (5.1) | 1010 (5.2) | 3758 (4.5) |
| Ischemic stroke | 414 (2.8) | 692 (7.6) | 3072 (7.6) | 1709 (8.7) | 5887 (7.1) |
| Intracranial hemorrhage | 50 (0.3) | 134 (1.5) | 611 (1.5) | 244 (1.2) | 1039 (1.2) |
| Osteoarthritis | 1769 (12.1) | 1619 (17.8) | 7418 (18.5) | 3675 (18.7) | 14,481 (17.3) |
| Depression | 10,137 (69.5) | 4693 (51.6) | 20,393 (50.7) | 9478 (48.3) | 44,701 (53.5) |

[a] Within 365 days after the cohort entry date
[b] HbA1c results were not considered for imputation because ~90% of those with HbA1c also had either a fasting or a random baseline glucose result
[c] Determined over the 183 days prior to the cohort entry date; Gagne et al (REF)
[d] Diabetes at baseline was an exclusion for Baseline Confounder Test Case 1 and therefore not included for this Test Case

**Table 6. Characteristics of Individuals in the Baseline Confounder Test Case 1 Population: Initiation of Second Generation Antipsychotics by Data Partner Site**

| Characteristics | Data Partner Site | | |
|---|---|---|---|
| | Site 1 N=3851 (4.6%) | Site 2 N=31,619 (37.9%) | Site 3 N=48,021 (57.5%) |
| Outcome: diabetes diagnosis after SGA initiation [a] | 170 (4.4) | 1515 (4.8) | 6268 (13.1) |
| Any blood glucose laboratory test result (fasting or random glucose, HbA1c) | 2232 (58.0) | 18,892 (59.8) | 13,301 (27.7) |
| Fasting glucose result available | 644 (16.7) | 7187 (22.7) | 170 (0.4) |
| Random glucose result available | 1856 (48.2) | 14,281 (45.2) | 13,032 (27.1) |
| HbA1c[b]result available | 135 (3.5) | 2516 (8.0) | 1865 (3.9) |
| Female sex | 2451 (63.6) | 19,639 (62.1) | 29,799 (62.1) |
| Age in years, mean (SD) | 55.6 (21.1) | 55.2 (21.1) | 64.2 (19.0) |
| Year of cohort entry | | | |
| 2008 | 479 (12.4) | 4437 (14.0) | 5934 (12.4) |
| 2009 | 878 (22.8) | 7363 (23.3) | 11,326 (23.6) |
| 2010 | 866 (22.5) | 7047 (22.3) | 10,715 (22.3) |
| 2011 | 909 (23.6) | 6859 (21.7) | 10,306 (21.5) |
| 2012 (1/1 – 10/31) | 719 (18.7) | 5913 (18.7) | 9740 (20.3) |
| Race | | | |
| White | 2878 (74.7) | 23,709 (75.0) | 33,104 (68.9) |
| Black | 142 (3.7) | 2715 (8.6) | 3410 (7.1) |
| Other | 88 (2.3) | 2501 (7.9) | 560 (1.2) |
| Unknown | 743 (19.3) | 2694 (8.5) | 10,947 (22.8) |
| Hispanic ethnicity (Y vs N or unknown) | 304 (7.9) | 3633 (11.5) | 858 (1.8) |
| Baseline procedure code (Glucose or HbA1c[b] CPT code) | 1825 (47.4) | 12,239 (38.7) | 20,761 (43.2) |
| Number of unique medication classes dispensed, mean (SD)[c] | 9.1 (4.7) | 9.3 (4.8) | 10.7 (5.0) |
| Number of ambulatory medical visits during baseline, mean (SD)[c] | 6.7 (6.5) | 11.1 (18.1) | 16.8 (17.8) |
| Emergency department visit during baseline, N (%) yes [c] | 1366 (35.5) | 14,693 (46.5) | 15,806 (32.9) |
| Hospitalization during baseline, N (%) yes [c] | 987 (25.6) | 8725 (27.6) | 12,468 (26.0) |
| Institutional stay during baseline, N (%) yes [c] | 372 (9.7) | 2823 (8.9) | 15,750 (32.8) |
| Comorbidity Score, mean (SD) [c,,d] | 1.9 (2.1) | 1.7 (2.0) | 2.1 (2.5) |
| Individual comorbidities [c,,d] | | | |
| Alcohol abuse | 444 (11.5) | 3517 (11.1) | 2709 (5.6) |
| Anemia, deficiency | 335 (8.7) | 3270 (10.3) | 8847 (18.4) |
| Arrhythmia | 401 (10.4) | 3265 (10.3) | 7798 (16.2) |
| Coagulation disorder | 102 (2.6) | 737 (2.3) | 1507 (3.1) |
| Congestive heart failure | 265 (6.9) | 2001 (6.3) | 6317 (13.2) |
| Diabetes, complicated [d] | NA | NA | NA |
| Dementia | 553 (14.4) | 2969 (9.4) | 1152 (24.1) |
| Fluid/Electrolyte disorders | 600 (15.6) | 3574 (11.3) | 8528 (17.8) |
| HIV/AIDS | 7 (0.2) | 142 (0.4) | 185 (0.4) |
| Hypertension | 1188 (30.8) | 10,800 (34.2) | 25,684 (53.5) |
| Hemiplegia | 37 (1.0) | 343 (1.1) | 980 (2.0) |
| Liver disease | 104 (2.7) | 911 (2.9) | 1185 (2.5) |
| Metastatic cancer | 31 (0.8) | 549 (1.7) | 708 (1.5) |
| Psychosis | 2831 (73.5) | 20,364 (64.4) | 23,434 (48.8) |
| Pulmonary disease, chronic | 597 (15.5) | 4900 (15.5) | 10,047 (20.9) |
| Pulmonary circulation disorder | 100 (2.6) | 363 (1.10 | 1036 (2.2) |
| Peripheral vascular disease | 176 (4.6) | 1823 (5.8) | 5781 (12.0) |
| Renal failure | 357 (9.3) | 2452 (7.8) | 4831 (10.1) |
| Tumor, any | 183 (4.8) | 1778 (5.6) | 3764 (7.8) |

| Characteristics | Data Partner Site | | |
|---|---|---|---|
| | Site 1 N=3851 (4.6%) | Site 2 N=31,619 (37.9%) | Site 3 N=48,021 (57.5%) |
| Weight loss | 120 (3.1) | 819 (2.6) | 1676 (3.5) |
| Additional comorbidities (specific to test case) | | | |
| Myocardial infarction | 138 (3.6) | 1311 (4.1) | 2309 (4.8) |
| Ischemic stroke | 141 (3.7) | 937 (3.0) | 4809 (10.0) |
| Intracranial hemorrhage | 48 (1.2) | 313 (1.0) | 678 (1.4) |
| Osteoarthritic | 501 (13.0) | 4112 (13.0) | 9868 (20.5) |
| Depression | 2309 (60.0) | 17983 (56.9) | 24409 (50.8) |

[a] Within 365 days after the cohort entry date
[b] HbA1c results were not considered for imputation because ~90% of those with HbA1c also had either a fasting or a random baseline glucose result
[c] Determined over the 183 days prior to the cohort entry date; Gagne et al[69]
[d] Diabetes at baseline was an exclusion for Baseline Confounder Test Case 1 and therefore not included for this Test Case

### c. Baseline Confounder Test Case 1 methods to investigate predictors of missing glucose results values

We fit logistic regression models to investigate variables predictive of missing glucose values and how that varied by site. In these models, the outcome was missing glucose (yes/no). Predictors included all variables shown in **Table 6**. This modeling was done overall (pooled) and separately by site. The results are presented in Table 7. In general, in these multivariable fully-adjusted models, the odds of missing glucose were higher for subjects taking aripiprazole relative to other drugs. The odds of missingness decreased from 2008 to 2012. For many variables, the ORs were similar across sites. However, exceptions such as sex (OR men vs women less than 1 in site 2 and greater than 1 in site 3), Hispanic ethnicity and intracranial hemorrhage (large OR in site 1 and small OR in site 2) were noted.

**Table 7. Baseline Confounder Test Case 1, Initiation of Second Generation Antipsychotics and Missingness of Glucose Results: Adjusted Logistic Regression Models assessing Associations with Missing Baseline Glucose Results Overall and by Data Partner Site**

| Characteristic [a] | Outcome Missing Glucose (1=missing glucose, 0=has glucose laboratory test result) Adjusted Odds Ratio (95% CI) | | | |
|---|---|---|---|---|
| | All Data Partner Sites Combined | Data Partner Site | | |
| | | Site 1 | Site 2 | Site 3 |
| SGA, Aripiprazole reference | | | | |
| Olanzapine | 0.81 (0.76, 0.85) | 0.63 (0.44, 0.91) | 0.83 (0.75, 0.93) | 0.99 (0.91, 1.07) |
| Quetiapine | 0.77 (0.73, 0.80) | 0.74 (0.58, 0.94) | 0.93 (0.87, 1.00) | 0.89 (0.84, 0.95) |
| Risperidone | 0.85 (0.81, 0.90) | 0.70 (0.55, 0.89) | 0.94 (0.86, 1.02) | 0.83 (0.77, 0.89) |
| Sex, male vs. female | 1.03 (0.99, 1.06) | 0.93 (0.79, 1.09) | 0.91 (0.86, 0.96) | 1.07 (1.02, 1.12) |
| Age (per 10 years) | 0.95 (0.94, 0.96) | 0.87 (0.82, 0.91) | 0.86 (0.85, 0.88) | 0.99 (0.97, 1.01) |
| Outcome: diabetes diagnosis after SGA initiation | 1.16 (1.08, 1.25) | 1.05 (0.62, 1.78) | 0.89 (0.74, 1.08) | 0.86 (0.79, 0.93) |
| Year of cohort entry, 2008 reference | | | | |
| 2009 | 0.97 (0.93, 1.03) | 0.78 (0.61, 1.01) | 0.90 (0.82, 0.98) | 1.06 (0.98, 1.14) |
| 2010 | 0.85 (0.81, 0.90) | 0.65 (0.51, 0.84) | 0.88 (0.80, 0.97) | 0.86 (0.79, 0.92) |
| 2011 | 0.79 (0.75, 0.83) | 0.52 (0.40, 0.67) | 0.83 (0.76, 0.91) | 0.77 (0.71, 0.83) |
| 2012 | 0.77 (0.73, 0.81) | 0.56 (0.43, 0.73) | 0.86 (0.79, 0.95) | 0.67 (0.62, 0.73) |
| Hispanic ethnicity (Y vs N or unknown) | 0.48 (0.45, 0.52) | 0.82 (0.61, 1.11) | 1.00 (0.91, 1.10) | 0.48 (0.41, 0.56) |
| Race, unknown reference | | | | |
| White | 0.76 (0.72, 0.80) | 0.77 (0.63, 0.95) | 1.08 (0.97, 1.21) | 1.20 (1.12, 1.28) |

| Characteristic [a] | Outcome Missing Glucose (1=missing glucose, 0=has glucose laboratory test result) Adjusted Odds Ratio (95% CI) | | | |
|---|---|---|---|---|
| | All Data Partner Sites Combined | Data Partner Site | | |
| | | Site 1 | Site 2 | Site 3 |
| Black | 0.75 (0.70, 0.80) | 1.04 (0.68, 1.59) | 1.21 (1.05, 1.39) | 1.12 (1.02, 1.24) |
| Other | 0.47 (0.43, 0.51) | 1.02 (0.61, 1.69) | 0.90 (0.78, 1.04) | 1.33 (1.08, 1.64) |
| Number of baseline medication classes | 0.97 (0.96, 0.97) | 0.94 (0.92, 0.96) | 0.97 (0.96, 0.98) | 0.96 (0.95, 0.96) |
| Number of ambulatory medical visits | 1.01 (1.01, 1.01) | 0.97 (0.96, 0.99) | 0.99 (0.99, 0.99) | 1.00 (1.00, 1.00) |
| Emergency department visits (Y vs N) | 0.77 (0.76, 0.78) | 0.81 (0.74, 0.89) | 0.43 (0.41, 0.45) | 1.08 (1.05, 1.10) |
| Hospitalization (Y vs N) | 0.76 (0.74, 0.78) | 0.99 (0.84, 1.15) | 0.48 (0.44, 0.52) | 1.14 (1.10, 1.18) |
| Institutional stays (Y vs N) | 1.42 (1.39, 1.45) | 0.49 (0.35, 0.69) | 0.80 (0.68, 0.94) | 1.11 (1.08, 1.13) |
| Gagne et al individual comorbidities | | | | |
| Alcohol abuse | 0.67 (0.63, 0.71) | 1.16 (0.91, 1.48) | 0.97 (0.88, 1.07) | 0.97 (0.88, 1.07) |
| Anemia, deficiency | 0.79 (0.76, 0.83) | 0.90 (0.62, 1.30) | 0.64 (0.55, 0.73) | 0.69 (0.65, 0.73) |
| Arrhythmia | 0.92 (0.87, 0.96) | 1.08 (0.77, 1.50) | 0.74 (0.65, 0.86) | 0.97 (0.91, 1.03) |
| Coagulation disorder | 0.79 (0.72, 0.87) | 0.58 (0.27, 1.23) | 0.48 (0.33, 0.70) | 0.92 (0.82, 1.04) |
| Congestive heart failure | 1.19 (1.12, 1.26) | 0.72 (0.44, 1.17) | 1.05 (0.86, 1.28) | 1.07 (1.00, 1.15) |
| Dementia | 1.24 (1.19, 1.30) | 0.80 (0.61, 1.05) | 0.95 (0.84, 1.06) | 1.06 (1.00, 1.12) |
| Fluid/Electrolyte disorders | 1.00 (0.95, 1.05) | 0.44 (0.32, 0.59) | 0.36 (0.30, 0.44) | 1.10 (1.03, 1.17) |
| HIV/AIDS | 0.35 (0.27, 0.44) | 0.00 (0.00, ) | 0.39 (0.25, 0.62) | 0.27 (0.20, 0.37) |
| Hypertension | 0.77 (0.74, 0.80) | 0.60 (0.49, 0.74 ) | 0.61 (0.56, 0.65) | 0.68 (0.65, 0.72 ) |
| Hemiplegia | 1.02 (0.90, 1.16) | 1.31 (0.55, 3.14) | 0.78 (0.49, 1.26) | 1.14 (0.98, 1.34) |
| Liver disease | 0.66 (0.60, 0.73) | 0.74 (0.42, 1.29) | 0.66 (0.54, 0.80) | 0.62 (0.55, 0.70) |
| Metastatic cancer | 0.66 (0.58, 0.76) | 0.91 (0.23, 3.61) | 1.02 (0.73, 1.42 ) | 0.86 (0.72, 1.03) |
| Psychosis | 0.80 (0.77, 0.83) | 0.89 (0.73, 1.09) | 0.83 (0.78, 0.89 ) | 0.87 (0.83, 0.91) |
| Pulmonary disease, chronic | 0.99 (0.95, 1.03) | 0.79 (0.63, 1.00) | 0.87 (0.79, 0.95) | 0.93 (0.88, 0.98) |
| Pulmonary circulation disorder | 1.08 (0.96, 1.22) | 0.83 (0.39, 1.76) | 1.02 (0.59, 1.76) | 1.14 (0.99, 1.32) |
| Peripheral vascular disease | 0.96 (0.91, 1.01) | 0.87 (0.51, 1.48) | 0.88 (0.73, 1.05) | 0.86 (0.81, 0.92) |
| Renal failure | 0.49 (0.47, 0.52) | 0.66 (0.46, 0.96) | 0.81 (0.70, 0.93) | 0.46 (0.43, 0.49) |
| Tumor, any | 0.71 (0.66, 0.75) | 0.63 (0.39, 1.02) | 0.74 (0.62, 0.88) | 0.69 (0.63, 0.74) |
| Weight loss | 0.95 (0.86, 1.04) | 1.16 (0.61, 2.20) | 0.70 (0.46, 1.08) | 1.19 (1.06, 1.35) |
| Additional comorbidities specific to test case | | | | |
| Myocardial infarction | 0.81 (0.75, 0.87) | 0.82 (0.46, 1.46) | 0.82 (0.66, 1.03) | 0.85 (0.78, 0.94) |
| Ischemic stroke | 1.23 (1.15, 1.31) | 1.32 (0.75, 2.31) | 0.78 (0.58, 1.05) | 0.94 (0.87, 1.01) |
| Intracranial hemorrhage | 0.86 (0.75, 0.99) | 2.51 (1.12, 5.64) | 0.38 (0.18 , 0.82) | 1.10 (0.91, 1.32) |
| Osteoarthritis | 0.91 (0.88, 0.95) | 0.85 (0.65, 1.12) | 0.77 (0.69 , 0.85) | 0.95 (0.90, 1.01) |
| Depression | 0.90 (0.87, 0.93) | 1.10 (0.93, 1.31) | 0.94 (0.88, 1.00) | 0.89 (0.85, 0.93) |

[a] CPT Code for glucose not included due to zero cell/collinearity; time to outcome was also included in these models (to align with variables in the imputation models) but was not included in the table due to limited interpretability and non-significance

### d. Baseline Confounder Test Case 1 analytic approaches to handle missing glucose result values: multiple imputation using regression, multiple Imputation using predictive mean matching, and meta-analysis

We compared five different methods (Table 8) to assess the importance of the baseline glucose result value in estimating the relationship of SGA with diabetes risk. Three (naïve) methods were used to "avoid" imputation of missing data (Table 8, Models 1 – 3). The first approach used the entire cohort, but did not include glucose in the outcome model. In other words, this approach was to not use the laboratory result variable regardless of whether the value was present or missing. This approach would be valid if there was no residual confounding due to the baseline glucose results value (i.e., glucose could appear as a confounder in bivariate analysis, but not after controlling for demographics, utilization, diagnoses, and other information). The second and third approaches involved excluding all data from subjects who had missing glucose results values (i.e., complete case analyses). This approach is not recommended, but was used for comparison. In this complete case cohort, we fit one outcome model that controlled for all variables except glucose (Table 8, Model 2). In the other, we included all variables (Table 8, Model 3).

After completing the above models, we applied two MI methods (Table 8, Models 4 and 5): a regression approach and predictive mean matching. In each case, we imputed ten data sets and combined the results using the methods of Rubin.[70] Both the regression method and predictive mean matching were implemented using SAS® PROC MI (SAS Institute Inc., Cary, NC).[71]

The regression approach to MI is the most common and easy to implement MI method.[70] Because it relies on a normality assumption, we first assessed skewness and performed a log-transformation to improve symmetry, with log glucose serving as the outcome variable in the imputation. All of the variables listed in **Table 6** were included in the model to predict log-glucose. A linear regression was fitted with log-glucose as the outcome and the other variables as predictors for subjects with observed glucose. Then, for each patient missing log-glucose, a draw from the predictive distribution was obtained. This created one imputed data set. The process was repeated nine more times.

Predictive mean matching is similar, except the draw from the predictive distribution is not used as the imputed value.[72] Instead, the predicted value is matched to a randomly selected observed value of log-glucose similar to the predicted value. An advantage of predictive mean matching is that normality is not assumed. In addition, imputed values are guaranteed to not be more extreme than observed values. A drawback of the approach is the tradeoff between having matched values close to the predicted values, and having low between imputation correlations. If, for example, the number of 'similar' values were small, then the correlation between imputed data sets would be high. Alternatively, having a more liberal definition can lead to less correlation with the predicted value (i.e., more random noise). Although not implemented in this test case, one way to deal with this problem is to do weighted sampling which lowers the correlation by distributing the matching to more than one 'best' observation.[73, 74]

The outcome model was a Cox proportional hazards model. The SGAs were the exposures of interest, and were entered into the model as separate indicator variables with aripiprazole as the reference. Log glucose was included in the models (except for the two models) as a potential confounder. All other variables in **Table 6** were included in the model.

We made one additional comparison. We compared results based on pooling the data from all three sites and treating it as one big data set (with site indicators in the model) versus imputing and fitting the

outcome model separately each site and then combining using meta-analysis methods. For the separate site analysis, we used the same imputation methods and outcome model. For the meta-analysis, we used a Bayesian random effects approach (implemented using PROC MCMC in SAS).

### e. Baseline Confounder Test Case 1 results and discussion

The adjusted results from the Cox models are in Table 8, with hazard ratios (HR) and 95%CI reported for the SGA and for log-glucose. HRs for other variables are not reported (to focus attention on the main findings).

**Table 8. Assessment of whether Risk of Diabetes Varies by Specific Second Generation Antipsychotic Agent. Hazard Ratios from Statistical Models with and without Imputation of Baseline Glucose [a,b]**

| Characteristic | Statistical Model Hazard Ratios (95% CI) | | | | |
|---|---|---|---|---|---|
| | 1 Entire Cohort (with and without baseline glucose results); Glucose Results Not Included | 2 Sub-Cohort (Complete Cases) with Baseline Glucose Results Available; Glucose Results Not Included | 3 Sub-Cohort (Complete Cases) with Baseline Glucose Results Available; Glucose Results Included | 4 Entire Cohort; Baseline Glucose Results Included and Missing Baseline Glucose Results Imputed using Predictive Mean Matching | 5 Entire Cohort; Baseline Glucose Results Included and Missing Baseline Glucose Results Imputed using Regression |
| **Imputation across data partner sites** | | | | | |
| N in model | 83,491 | 33,714 | 33,714 | 83,491 | 83,491 |
| Olanzapine | 1.04 (0.92, 1.18) | 0.92 (0.74, 1.14) | 0.91 (0.73, 1.12) | 1.02 (0.90, 1.16) | 1.02 (0.91, 1.16) |
| Quetiapine | 1.04 (0.9 , 1.15) | 1.05 (0.90, 1.23) | 1.04 (0.89, 1.21) | 1.03 (0.93, 1.13) | 1.02 (0.93, 1.12) |
| Risperidone | 1.11 (1.00, 1.23) | 1.17 (0.99, 1.38) | 1.17 (0.99, 1.38) | 1.10 (0.99, 1.22) | 1.10 (0.99, 1.22) |
| Log GLU | NA | NA | 12.8 (10.2, 16.2) | 12.2 (10.1, 14.7) | 16.5 (12.8, 21.2) |
| **Imputation within each data partner site; outcome model results combined with meta-analysis** | | | | | |
| N in model | 83,491 | NA [c] | NA [c] | 83,491 | 83,491 |
| Olanzapine | 0.99 (0.41, 1.75) | | | 0.97 (0.39, 1.76) | 0.96 (0.38, 1.76) |
| Quetiapine | 1.01 (0.57, 1.57) | | | 1.00 (0.56, 1.56) | 1.00 (0.55, 1.55) |
| Risperidone | 1.14 (0.70, 1.73) | | | 1.11 (0.65, 1.74) | 1.10 (0.64, 1.72) |
| **Imputation within each data partner site** | | | | | |
| Site 1 | | | | | |
| N in model | 3851 | 2193 | 2193 | 3851 | 3851 |
| Olanzapine | 0.62 (0.16, 2.39) | 0.32 (0.04, 2.89) | 0.27 (0.03, 2.46) | 0.57 (0.15, 2.19) | 0.53 (0.14, 2.04) |
| Quetiapine | 0.90 (0.42, 1.95) | 1.15 (0.42, 3.17) | 1.11 (0.40, 3.08) | 0.86 (0.40, 1.87) | 0.83 (0.38, 1.81) |
| Risperidone | 1.26 (0.59, 2.69) | 1.40 (0.50, 3.90) | 1.28 (0.45, 3.62) | 1.19 (0.55, 2.56) | 1.14 (0.53, 2.44) |
| Log GLU | | | 11.0 (3.54, 34.0) | 9.73 (2.83, 33.5) | 15.0 (5.16, 43.4) |
| Site 2 | | | | | |
| N in model | 31,619 | 18,530 | 18,530 | 31,619 | 31,619 |
| Olanzapine | 0.76 (0.56, 1.03) | 0.81 (0.53, 1.23) | 0.76 (0.50, 1.16) | 0.74 (0.54, 1.02) | 0.73 (0.54, 1.00) |
| Quetiapine | 0.85 (0.69, 1.06) | 1.12 (0.82, 1.53) | 1.07 (0.78, 1.46) | 0.83 (0.67, 1.04) | 0.82 (0.67, 1.02) |
| Risperidone | 1.01 (0.79, 1.29) | 1.27 (0.89, 1.80) | 1.10 (0.77, 1.56) | 0.92 (0.71, 1.18) | 0.90 (0.70, 1.15) |
| Log GLU | | | 44.7 (31.4, 63.7) | 41.8 (28.9, 60.3) | 62.8 (43.3, 90.9) |
| Site 3 | | | | | |
| N in model | 48,021 | 12,991 | 12,991 | 48,021 | 48,021 |
| Olanzapine | 1.13 (0.99, 1.29) | 1.00 (0.78, 1.27) | 0.99 (0.77, 1.27) | 1.13 (0.98, 1.29) | 1.12 (0.98, 1.28) |
| Quetiapine | 1.10 (0.99, 1.22) | 1.02 (0.85, 1.23) | 1.01 (0.84, 1.22) | 1.09 (0.98, 1.21) | 1.09 (0.98, 1.21) |
| Risperidone | 1.13 (1.01, 1.27) | 1.11 (0.91, 1.35) | 1.12 (0.92, 1.37) | 1.15 (1.02, 1.29) | 1.15 (1.02, 1.29) |
| Log GLU | | | 6.26 (4.59, 8.55) | 5.55 (4.14, 7.44) | 6.80 (4.84, 9.55) |

| Characteristic | Statistical Model Hazard Ratios (95% CI) | | | | |
|---|---|---|---|---|---|
| | 1 Entire Cohort (with and without baseline glucose results); Glucose Results Not Included | 2 Sub-Cohort (Complete Cases) with Baseline Glucose Results Available; Glucose Results Not Included | 3 Sub-Cohort (Complete Cases) with Baseline Glucose Results Available; Glucose Results Included | 4 Entire Cohort; Baseline Glucose Results Included and Missing Baseline Glucose Results Imputed using Predictive Mean Matching | 5 Entire Cohort; Baseline Glucose Results Included and Missing Baseline Glucose Results Imputed using Regression |
| [a] All variables shown in **Table 6** were included in the models<br>[b] Aripiprazole was reference SGA; site 1 was reference Data Partner<br>[c] Meta-analyses combining the results were completed for the primary Models 1, 4, and 5 | | | | | |

We first focus on the pooled analysis where data from sites were combined and then analyzed. These are displayed in the first rows of Table 8. As expected, log-glucose was a strong predictor of incident diabetes, with estimated HRs ranging from about 12 to 16 (and highly significant). The two MI methods (Table 8, Models 4 and 5) led to very similar point estimates and 95%CI for the SGAs, with each point estimate and confidence limit not differing by more than 0.01. Including log-glucose in the model had a small impact on the point estimates and 95%CI (Table 8, compare Model 1 with 4 and 5), but the difference was not clinically meaningful. The two models using only subjects with observed glucose (Table 8, Models 2 and 3) resulted in somewhat different HR and much wider 95%CI.

Next, we compare the pooled approach with the meta-analysis approach. We will focus on the MI approaches (Table 8 columns 4 and 5). First, notice the site-specific HR using predictive mean matching for imputation differ markedly across sites. For example, the estimated HR for olanzapine is 0.57 for site 1 and 1.13 for site 3. The 95%CI are also wide. Comparing the meta-analysis results with the pooled results shows drastic differences, especially in the width of the 95%CI (with the meta-analysis intervals being much wider). This suggests that pooling data and only adjusting for site indicators did not sufficiently account for variations in the relationship between variables between sites.

In summary, in this test case all three missing data methods led to similar point estimates and 95%CI. The most striking finding from this test case was how different the results were with a single pooled analysis rather than either site-specific or meta-analytic approach.

## 2. Baseline Confounder Test Case 2: Baseline serum creatinine in patients with diabetes starting an angiotensin converting enzyme inhibitor (ACEi) and risk of hyperkalemia

### a. Baseline Confounder Test Case 2 cohort development

ACEi are prescribed to selected persons with diabetes to lower blood pressure, slow renal disease progression, and decrease morbidity and mortality after myocardial infarction.[75-79] A known risk with ACEi is the potential for hyperkalemia.[80-83] Risk of this adverse effect is heightened for persons with poor kidney function or taking higher doses.[83, 84] This test case examined models that included adjustment for serum creatinine result values in analyses comparing high versus low dose lisinopril and the outcome of diagnosed hyperkalemia. As with all test cases in this report, serum creatinine result values were not consistently available and the primary goal was to describe methods of dealing with the missing laboratory results data. Additional detail on the research questions and the cohort are provided below.

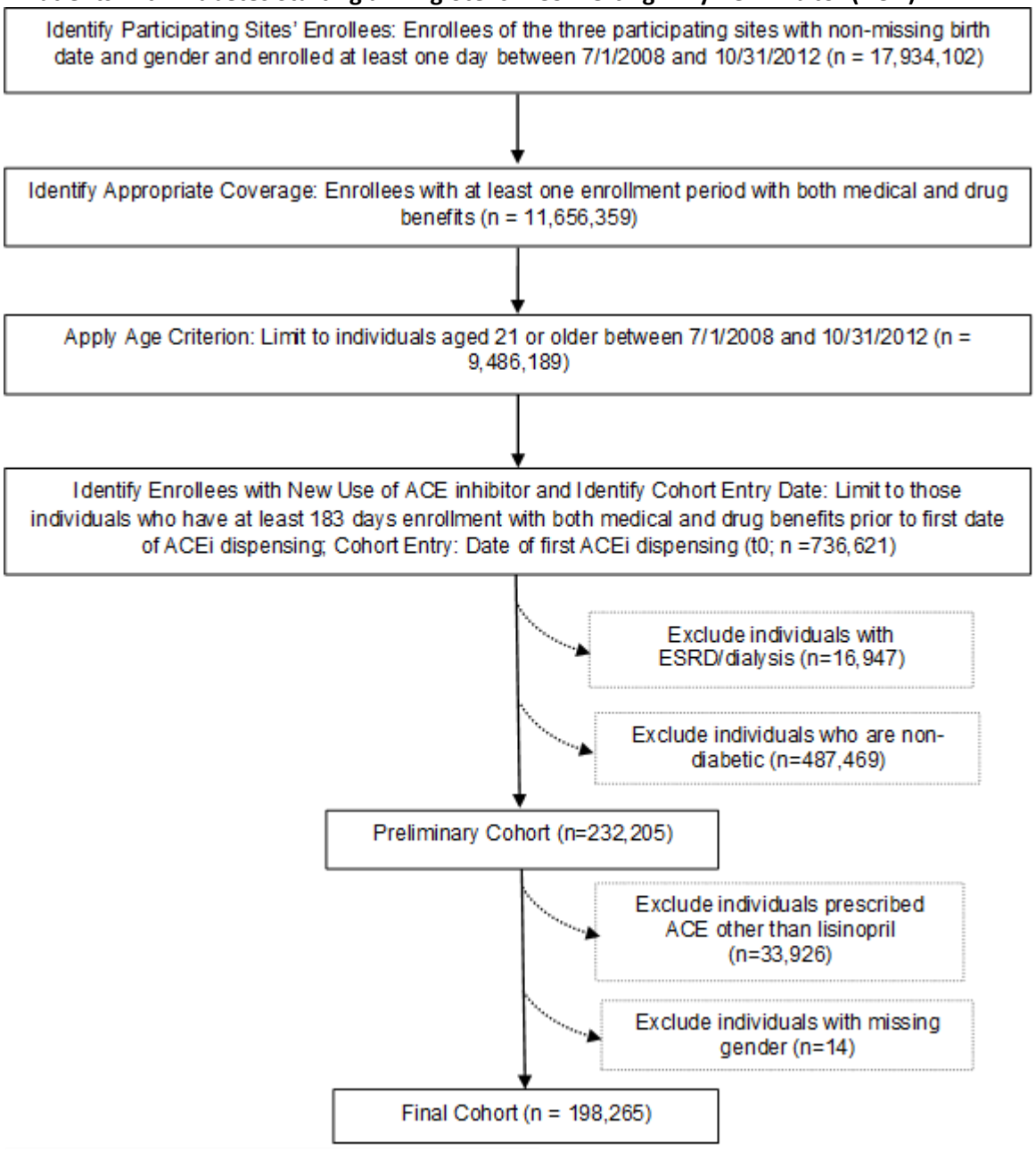
- Research Questions:

- Does inclusion of a baseline serum creatinine laboratory test result value reduce the potential bias when modeling the associations between exposure to lisinopril and the risk of a hyperkalemia outcome in patients in the MSDD with diagnosed diabetes?
- What is the contribution of a baseline serum creatinine laboratory result value when evaluating the association between high dose (≥ 20 mg) versus low dose (< 20 mg) lisinopril exposure and the risk of hyperkalemia outcome in patients in the MSDD with diagnosed diabetes? (serum creatinine result value is modeled as a continuous variable)
- What is the performance of selected "Missing Data Methods" when modeling the association between high versus low dose lisinopril exposure and hyperkalemia outcome?

Cohort members were adults ≥21 years of age with diagnosed diabetes who were members of one of the sites between 7/1/2008 and 10/31/2012. We identified new users of an ACEi by selecting the first dispensing and requiring a minimum of 6 months enrollment prior to the first dispensing. We excluded persons on dialysis or with end stage renal disease because they typically have rapidly changing clinical status and dialysis patients could have potassium values that rise and fall relative to dialysis timing. We retained patients with lesser degrees of chronic kidney disease. Serum creatinine results were identified and used only from outpatient care settings.

Lisinopril was the predominant ACEi used at all three data partner sites. We focused on lisinopril, retaining 85% of the cohort of new ACEi users (Figure 2).

**Figure 2. Test Case Cohort for Baseline Confounder Adjustment Test Case 2, Baseline Serum Creatinine in Patients with Diabetes Starting an Angiotensin Converting Enzyme Inhibitor (ACEi) [a]**




[a] Definitions and codes are in Appendix D

### b. Baseline Confounder Test Case 2 descriptive analysis of missing laboratory results

A comprehensive description of the cohort is provided in Tables 9 and 10. At the time of lisinopril initiation, 69,760 (35.2%) were started on higher doses (≥ 20 mg). Approximately 2.5% had a hyperkalemia diagnosis within one year after lisinopril initiation; this proportion was similar for low and high doses (2.6% and 2.4%, respectively). Creatinine results were available for 55% of the cohort. A lower proportion of patients starting higher dosage lisinopril had a creatinine result available (49% vs 58% on lower doses). This difference was partially related to differences in site prescribing patterns and creatinine result availability: site 3 had a higher proportion of high dose patients (76.9% vs 62.1% low dose patients) and was more likely to have missing creatinine results (44.5% with results versus >76% at the sites 1 and 2 [Table 10]). Although the availability of creatinine results differed by site, CPT codes indicating a test had been completed were comparable (79.7% site 1, 76.5% site 2, 80.6% site 3).

**Table 9. Characteristics of Individuals in the Baseline Confounder Test Case 2 Population, Lisinopril Dose and Hyperkalemia Diagnosis**

| Characteristics | Lisinopril | | Total N=198,265 |
|---|---|---|---|
| | Dose < 20 mg N=128,505 (64.8%) | Dose ≥ 20 mg N=69,760 (35.2%) | |
| Outcome: Hyperkalemia diagnosis within one year[a] | 3325 (2.6) | 1681 (2.4) | 5006 (2.5) |
| Serum creatinine laboratory test outpatient results available | 74,566 (58.0) | 34,487 (49.4) | 109,053 (55.0) |
| Serum creatinine (mg/dl), mean (SD) | 0.97 (0.42) | 1.03 (0.51) | 0.99 (0.45) |
| Age in years, mean (SD) | 64.6 (13.6) | 65.9 (11.9) | 65.0 (13.1) |
| Race | | | |
| White | 80,422 (62.6) | 42,064 (60.3) | 122,486 (61.8) |
| Black | 13,698 (10.7) | 11,794 (16.9) | 25,492 (12.9) |
| Other | 11,040 (8.6) | 4064 (5.8) | 15,104 (7.6) |
| Unknown | 23,345 (18.2) | 11,838 (17.0) | 35,183 (17.7) |
| Hispanic ethnicity (Y vs N or unknown) | 13,089 (10.2) | 4668 (6.7) | 17,757 (9.0) |
| Female gender | 61332 (47.7) | 33699 (48.3) | 95031 (47.9) |
| Year of cohort entry | | | |
| 2008 | 15,739 (12.2) | 8675 (12.4) | 24,414 (12.3) |
| 2009 | 31,431 (24.5) | 18,668 (26.8) | 50,099 (25.3) |
| 2010 | 28,369 (22.1) | 16,885 (24.2) | 45,254 (22.8) |
| 2011 | 28,606 (22.3) | 14,618 (21.0) | 43,244 (21.8) |
| 2012 (1/1 – 10/31) | 24,360 (19.0) | 10,914 (15.6) | 35,274 (17.8) |
| Site | | | |
| 1 | 6660 (5.2) | 1837 (2.6) | 8497 (4.3) |
| 2 | 42,009 (32.7) | 14,257 (20.4) | 56,266 (28.4) |
| 3 | 79,836 (62.1) | 53,666 (76.9) | 133,502 (67.3) |
| Serum creatinine laboratory procedure code (CPT code) [b, c] | 103,154 (80.3) | 54,289 (77.8) | 157,443 (79.4) |
| Number of unique medication classes dispensed, mean (SD)[c] | 6.2 (4.7) | 6.1 (4.5) | 6.2 (4.7) |
| Number of ambulatory medical visits during baseline, mean (SD)[c] | 7.1 (7.7) | 6.9 (7.3) | 7.0 (7.6) |
| Emergency department visit during baseline, N (%) yes [c] | 19.1 | 16.2 | 18.1 |
| Hospitalization during baseline, N (%) yes [c] | 16.9 | 12.7 | 15.4 |
| Institutional stay during baseline, N (%) yes [c] | 8.7 | 8.4 | 8.6 |
| Comorbidity score, mean (SD) [c, d] | 1.1 (2.2) | 0.81 (2.0) | 1.0 (2.1) |

| Characteristics | Lisinopril | | Total N=198,265 |
|---|---|---|---|
| | Dose < 20 mg N=128,505 (64.8%) | Dose ≥ 20 mg N=69,760 (35.2%) | |
| Individual comorbidities [c, d] | | | |
| Alcohol abuse | 1831 (1.4) | 818 (1.2) | 2649 (1.3) |
| Anemia, deficiency | 14,695 (11.4) | 7962 (11.4) | 22,657 (11.4) |
| Arrhythmia | 1,4692 (11.4) | 6072 (8.7) | 20,764 (10.5) |
| Coagulation disorder | 2701 (2.1) | 1073 2.1 | 3774 (1.9) |
| Congestive heart failure | 5311 (4.1) | 2643 (3.8) | 7954 (4.0) |
| Diabetes, complicated | 42 (<0.1) | 16 (<0.1) | 58 (<0.1) |
| Dementia | 1524 (1.2) | 626 (0.9) | 2150 (1.1) |
| Fluid/Electrolyte disorders | < 6 (0) | < 6 (0) | < 6 (0) |
| HIV/AIDS | 190 (0.1) | 78 (0.1) | 268 (0.1) |
| Hypertension | 79,187 (61.6) | 54,303 (77.8) | 133,490 (67.3) |
| Hemiplegia | < 6 (0) | < 6 (0) | < 6 (0) |
| Liver disease | 3716 (2.9) | 1553 (2.2) | 5269 (2.7) |
| Metastatic cancer | < 6 (0) | < 6(0) | < 6 (0) |
| Psychosis | 31 (<0.1) | 14 (<0.1) | 45 (<0.1) |
| Pulmonary disease, chronic | 13,091 (10.2) | 6188 (8.9) | 19,279 (9.7) |
| Pulmonary circulation disorder | 24 (<0.1) | 11 (<0.1) | 35 (<0.1) |
| Peripheral vascular disease | 8858 (6.9) | 4751 (6.8) | 13,609 (6.9) |
| Renal failure | 1505 (1.2) | 927 (1.3) | 2432 (1.2) |
| Tumor, any | 3442 (2.7) | 1893 (2.7) | 5335 (2.7) |
| Weight loss | 191 (0.1) | 65 (0.1) | 256 (0.1) |
| Additional comorbidities (specific to test case) | | | |
| Prior history of Hyperkalemia dx | 1500 (1.2) | 759 (1.1) | 2259 (1.1) |
| CKD I-IV | 17,918 (13.9) | 9751 (14.0) | 27,669 (14.0) |
| MI / Stroke | 16,722 (13.0) | 7530 (10.8) | 24,252 (12.2) |
| Rx dispensing, increases $K^+$ [e] | 63,207 (49.2) | 35,687 (51.2) | 98,894 (49.9) |
| Rx dispensing, decreases $K^+$ [e] | 7136 (5.6) | 2867 (4.1) | 1003 (5.0) |

[a] Within 365 days after the cohort entry date
[b] CPT codes 80047, 80048, 80053, 80069, 82565
[c] Determined over the 183 days prior to the cohort entry date
[d] Gagne et al[69]
[e] Dispensed in 100 days prior to the cohort entry date; medication list is in Appendix D

**Table 10. Characteristics of Individuals in the Baseline Confounder Test Case 2 Population, Lisinopril Dose and Hyperkalemia Diagnosis by Data Partner Site**

| Characteristics | Data Partner Site | | |
|---|---|---|---|
| | Site 1 N=8497 (4.3%) | Site 2 N=56,266 (28.4%) | Site 3 N=133,502 (67.3%) |
| Outcome: Hyperkalemia diagnosis within one year [a] | 168 (2.0) | 1052 (1.9) | 3786 (2.8) |
| Serum creatinine laboratory outpatient results available | 6716 (79.0) | 42,920 (76.3) | 59,417 (44.5) |
| Serum creatinine (mg/dl), mean (SD) | 1.00 (0.39) | 0.93 (0.47) | 1.04 (0.44) |
| Age in years, mean (SD) | 59.7 (13.0) | 58.7 (13.7) | 68.1 (11.7) |
| Race | | | |
| White | 4998 (58.8) | 28,470 (50.6) | 89,018 (66.7) |
| Black | 549 (6.5) | 6057 (10.8) | 18,886 (14.1) |
| Other | 391 (4.6) | 11,349 (20.2) | 3364 (2.5) |
| Unknown | 2559 (30.1) | 10,390 (18.5) | 22,234 (16.7) |
| Hispanic ethnicity (Y vs N or unknown) | 1812 (21.3) | 13,019 (23.1) | 2926 (2.2) |
| Female Gender | 3875 (45.6) | 24,891 (44.2) | 66,265 (49.6) |

| Characteristics | Data Partner Site | | |
|---|---|---|---|
| | Site 1 N=8497 (4.3%) | Site 2 N=56,266 (28.4%) | Site 3 N=133,502 (67.3%) |
| Year of Cohort Entry | | | |
| 2008 | 1058 (12.5) | 8226 (14.6) | 15,130 (11.3) |
| 2009 | 2191 (25.8) | 15,614 (27.8) | 32,294 (24.2) |
| 2010 | 1846 (21.7) | 12,988 (23.1) | 30,420 (22.8) |
| 2011 | 1986 (23.4) | 10,697 (19.0) | 30,541 (22.9) |
| 2012 (1/1 – 10/31) | 1416 (16.7) | 8741 (15.5) | 25,117 (18.8) |
| Serum creatinine laboratory Procedure Code (CPT code) [b, c] | 6775 (79.7) | 43,059 (76.5) | 107,609 (80.6) |
| Number of unique medication classes dispensed, mean (SD)[c] | 4.7 (4.0) | 4.8 (4.1) | 6.9 (4.8) |
| Number of ambulatory medical visits during baseline, mean (SD)[c] | 3.5 (4.2) | 4.1 (5.3) | 8.5 (8.1) |
| Emergency department visit during baseline, N (%) yes [c] | 1097 (12.9) | 11,959 (21.3) | 22,759 (17.0) |
| Hospitalization during baseline, N (%) yes [c] | 717 (8.4) | 6109 (10.9) | 23,795 (17.8) |
| Institutional stay during baseline, N (%) yes [c] | 135 (1.6) | 870 (1.5) | 16,005 (12.0) |
| Comorbidity Score, mean (SD) [c, d] | 0.8 (1.8) | 0.7 (1.8) | 1.2 (2.3) |
| Individual comorbidities [c, d] | | | |
| Alcohol abuse | 125 (1.5) | 983 (1.7) | 1541 (1.2) |
| Anemia, deficiency | 380 (4.5) | 3383 (6.0) | 18,894 (14.2) |
| Arrhythmia | 471 (5.5) | 3031 (5.4) | 17,262 (12.9) |
| Coagulation disorder | 120 (1.4) | 617 (1.1) | 3037 (2.3) |
| Congestive heart failure | 126 (1.5) | 439 (0.8) | 7389 (5.5) |
| Diabetes, complicated | < 6(0) | < 6 (0) | 56 (<0.1) |
| Dementia | 35 (0.4) | 232 (0.4) | 1883 (1.4) |
| Fluid/Electrolyte disorders | < 6 (0) | < 6 (0) | < 6 (0) |
| HIV/AIDS | < 6 (0) | 94 (0.2) | 171 (0.1) |
| Hypertension | 3635 (42.8) | 27,206 (48.4) | 102,649 (76.9) |
| Hemiplegia | < 6 (0) | < 6 (0) | < 6 (0) |
| Liver disease | 308 (3.6) | 1661 (3.0) | 3300 (2.5) |
| Metastatic cancer | < 6 (0) | < 6 (0) | < 6 (0) |
| Psychosis | < 6 (0) | < 6 (0) | 44 (<0.1) |
| Pulmonary disease, chronic | 383 (4.5) | 1794 (3.2) | 17,102 (12.8) |
| Pulmonary circulation disorder | < 6 (0) | < 6 (0) | 35 (<0.1) |
| Peripheral vascular disease | 187 (2.2) | 1462 (2.6) | 11,960 (9.0) |
| Renal failure | 59 (0.7) | 366 (0.7) | 2007 (1.5) |
| Tumor, any | 127 (1.5) | 952 (1.7) | 4256 (3.2) |
| Weight loss | < 6 (0) | 31 (0.1) | 223 (0.2) |
| Additional comorbidities (specific to test case) | | | |
| Prior history of hyperkalemia dx | 63 (0.7) | 366 (0.7) | 1830 (1.4) |
| CKD I-IV | 1223 (14.5) | 6095 (10.8) | 20,341 (15.2) |
| MI / Stroke | 542 (6.4) | 3943 (7.0) | 19,767 (14.8) |
| Rx dispensing, increases $K^+$ [e] | 3131 (36.8) | 23,190 (41.2) | 72,573 (54.4) |
| Rx dispensing, decreases $K^+$ [e] | 334 (3.9) | 1823 (3.2) | 7846 (5.9) |

[a] Within 365 days after the cohort entry date
[b] CPT codes 80047, 80048, 80053, 80069, 82565
[c] Determined over the 183 days prior to the cohort entry date
[d] Gagne et al[69]
[e] Dispensed in 100 days prior to the cohort entry date; medication list is in Appendix D

### c. Baseline Confounder Test Case 2 methods to investigate predictors of missing serum creatinine results value

Tables 9 and 10 provide both a summary comorbidity score[69] as well as indicators for the individual comorbidity items that are used in this score. It is often preferable to use the individual items when the goal is to maximize the amount of variability explained. MI modeling is one such arena; we chose to utilize the individual indicators rather than the summary score in the imputation models. However, a few individual indicators were very rare (≤0.1% overall) and were not used as individual items in any analytic models (complicated diabetes, fluid/electrolyte disorders, hemiplegia, metastatic cancer, psychosis, pulmonary circulation disorder). Some additional comorbidity items, not part of the comorbidity score, were also captured and for this test case: prior history of hyperkalemia, CKD, history of MI or stroke and indicators for use of other drugs that might impact serum potassium levels (list of these other drugs is in Appendix D).

We examined variable associations with missing creatinine results for the cohort overall (Table 11) and by site (Table 12). In these logistic regression models, missing creatinine result (indicator=1) versus non-missing creatinine result (indicator=0) was the outcome. In combined models that adjusted for site and other covariates (Table 11), persons on higher doses of lisinopril remained significantly more likely to have missing serum creatinine results but the magnitude of the ORs were decreased (OR 1.09 versus 1.41 unadjusted). The odds of creatinine being missing decreased each year after 2008; females were slightly less likely to have missing creatinine as were non-white race groups. A higher comorbidity score was associated with less missing creatinine results but histories of ED visits and hospitalizations both resulted in increased risk of missing creatinine results. A history of CKD was associated with decreased risk of missing creatinine results in both the unadjusted and adjusted models. Univariate and adjusted associations for additional covariates are in Table 11.

Site specific models in Table 12 highlight variability by site. For example, the pattern of less missing data for non-white race groups seen in the combined models is only at site 3. Similarly, although a CKD diagnosis is strongly associated with a decreased likelihood of missing creatinine results at all sites, the magnitude of the associations differs by site. Even variables with associations that appear similar may still have significant variability by site. For example, increased age is associated with decreased missing creatinine results and the associations are relatively consistent by site in unadjusted models (OR 0.88 to 0.94). Nevertheless, a site by age interaction term is significant ($p<0.001$). These general patterns support utilizing missing data methods that impute or account for missing data within site or use methods that allow for differing variable associations by site (e.g. site by variable interactions).

**Table 11. Characteristics Potentially Associated with Missing Baseline Serum Creatinine Laboratory Results for the Population Included in Baseline Confounder Test Case 2, Hyperkalemia Diagnosis after Lisinopril Initiation: Unadjusted and Adjusted Odds Ratios**

| Characteristic [b] | Odds Ratio (95% Confidence Interval) [a] | |
|---|---|---|
| | Unadjusted | Adjusted [c] |
| Lisinopril dose (≥ 20 mg vs < 20 mg) | 1.41 (1.39, 1.44) | 1.09 (1.06, 1.12) |
| Sex, female vs male | 0.98 (0.96, 1.00) | 0.93 (0.90, 0.95) |
| Age (per 10 years) | 1.09 (1.08, 1.10) | 0.96 (0.95, 0.97) |
| Race, White reference | | |
| Black | 0.88 (0.86, 0.91) | 0.75 (0.72, 0.77) |
| Other | 0.45 (0.43, 0.46) | 0.72 (0.67, 0.76) |
| Unknown | 0.84 (0.82, 0.86) | 0.66 (0.63, 0.69) |
| Hispanic ethnicity (Y vs N or unknown) | 0.39 (0.38, 0.41) | 0.77 (0.72, 0.83) |
| Year of Cohort Entry, 2008 reference | | |
| 2009 | 0.99 (0.96, 1.03) | 1.01 (0.96, 1.05) |
| 2010 | 0.84 (0.81, 0.86) | 0.79 (0.75, 0.82) |
| 2011 | 0.81 (0.78, 0.83) | 0.71 (0.68, 0.74) |
| 2012 (1/1 – 10/31) | 0.75 (0.72, 0.77) | 0.61 (0.59, 0.64) |
| Site, 3 reference | | |
| 1 | 0.21 (0.20, 0.22) | 0.04 (0.04, 0.04) |
| 2 | 0.25 (0.24, 0.26) | 0.04 (0.03, 0.04) |
| Serum Creatinine-related procedure code (CPT code) | 0.02 (0.02, 0.02) | 0.00 (0.00, 0.00) |
| Number of unique medication classes dispensed | 0.99 (0.99, 1.00) | 0.99 (0.99, 1.00) |
| Number of ambulatory visits | 1.01 (1.01, 1.01) | 1.01 (1.01, 1.01) |
| Emergency department visits (Y vs N) | 1.09 (1.08, 1.11) | 1.14 (1.12, 1.16) |
| Hospitalization (Y vs N) | 1.28 (1.26, 1.30) | 1.16 (1.13, 1.19) |
| Institutional stay (Y vs N) | 1.33 (1.31, 1.35) | 1.19 (1.17, 1.21) |
| Comorbidity score[d] | 0.93 (0.92, 0.93) | 0.91 (0.90, 0.92) |
| Prior hyperkalemia diagnosis (Y vs N) | 0.89 (0.82, 0.96) | 1.25 (1.13, 1.38) |
| CKD I-IV (Y vs N) | 0.39 (0.38, 0.41) | 0.47 (0.45, 0.49) |
| MI / Stroke (Y vs N) | 1.20 (1.17, 1.23) | 0.91 (0.88, 0.95) |
| Any Rx dispensing, increases $K^+$ [e] (Y vs N) | 1.09 (1.08, 1.11) | 1.07 (1.04, 1.10) |
| Any Rx dispensing, decreases $K^+$ [e] (Y vs N) | 1.06 (1.02, 1.11) | 1.03 (0.98, 1.09) |

[a] Missing baseline glucose laboratory test results is the dependent variable
[b] Determined over the 183 days prior to the cohort entry date
[c] Adjusted for all variables shown in table
[d] Gagne et al[69]
[e] Dispensed in 100 days prior to the cohort entry date; medication list is in Appendix D

**Table 12. Characteristics Potentially Associated with Missing Baseline Serum Creatinine Laboratory Results for the Population Included in Baseline Confounder Test Case 2, Hyperkalemia Diagnosis after Lisinopril Initiation: Unadjusted and Adjusted Odds Ratios by Data Partner Site**

| Characteristic [a] | Site 1 N=8,497 | | Site 2 N=56,266 | | Site 3 N=133,502 | |
|---|---|---|---|---|---|---|
| | Odds Ratio (95% Confidence Interval) [b] | | Odds Ratio (95% Confidence Interval) [b] | | Odds Ratio (95% Confidence Interval) [b] | |
| | Unadjusted | Adjusted [c] | Unadjusted | Adjusted [c] | Unadjusted | Adjusted [c] |
| Lisinopril Dose (≥ 20 mg vs < 20 mg) | 1.27 (1.12, 1.43) | 1.40 (0.94, 2.11) | 1.26 (1.21, 1.32) | 1.16 (0.97, 1.38) | 1.17 (1.14, 1.19) | 1.08 (1.05, 1.10) |
| Sex, Female vs Male | 0.93 (0.83, 1.03) | 1.20 (0.84, 1.73) | 0.92 (0.89, 0.96) | 1.05 (0.90, 1.24) | 0.91 (0.89, 0.93) | 0.92 (0.90, 0.94) |
| Age (per 10 years) | 0.92 (0.89, 0.96) | 1.12 (0.96, 1.30) | 0.95 (0.94, 0.97) | 1.00 (0.94, 1.06) | 0.90 (0.89, 0.91) | 0.95 (0.94, 0.96) |
| Race, White reference | | | | | | |
| Black | 1.03 (0.83, 1.29) | 1.52 (0.73, 3.18) | 1.17 (1.10, 1.25) | 1.09 (0.84, 1.41) | 0.79 (0.76, 0.81) | 0.75 (0.72, 0.77) |
| Other | 0.87 (0.66, 1.14) | 1.34 (0.51, 3.48) | 1.01 (0.96, 1.06) | 0.68 (0.55, 0.85) | 0.73 (0.68, 0.78) | 0.70 (0.64, 0.75) |
| Unknown | 1.42 (1.27, 1.59) | 3.86 (2.46, 6.06) | 1.09 (1.03, 1.15) | 0.66 (0.51, 0.86) | 0.89 (0.87, 0.92) | 0.65 (0.62, 0.67) |
| Hispanic ethnicity, (Y vs N or unknown) | 1.18 (1.04, 1.33) | 0.27 (0.16, 0.47) | 1.03 (0.98, 1.07) | 1.09 (0.87, 1.38) | 0.43 (0.39, 0.46) | 0.59 (0.54, 0.66) |
| Year of Cohort Entry, 2008 reference | | | | | | |
| 2009 | 0.86 (0.72, 1.02) | 0.77 (0.40, 1.51) | 1.00 (0.94, 1.06) | 1.70 (1.32, 2.19) | 0.94 (0.91, 0.98) | 0.98 (0.94, 1.03) |
| 2010 | 0.85 (0.70, 1.02) | 0.74 (0.37, 1.48) | 0.86 (0.81, 0.92) | 2.17 (1.67, 2.82) | 0.72 (0.69, 0.74) | 0.75 (0.72, 0.79) |
| 2011 | 1.04 (0.87, 1.24) | 1.76 (0.92, 3.35) | 0.83 (0.77, 0.89) | 2.10 (1.60, 2.77) | 0.64 (0.62, 0.67) | 0.67 (0.64, 0.70) |
| 2012 (1/1 – 10/31) | 0.96 (0.79, 1.16) | 1.72 (0.87, 3.40) | 0.76 (0.71, 0.82) | 2.27 (1.70, 3.04) | 0.59 (0.56, 0.61) | 0.58 (0.55, 0.61) |
| Serum creatinine-related procedure Code (CPT code)(Y vs N) | <0.001 (0.00, 0.00) | <0.001 (0.00, 0.00) | <0.001 (0.00, 0.00) | <0.001 (0.00, 0.00) | 0.04 (0.03, 0.04) | 0.04 (0.03, 0.04) |
| Number of unique medication classes dispensed | 0.94 (0.93, 0.95) | 0.77 (0.73, 0.82) | 0.92 (0.91, 0.92) | 0.90 (0.88, 0.92) | 0.97 (0.97, 0.98) | 0.99 (0.99, 1.00) |
| Number of ambulatory visits | 0.96 (0.95, 0.98) | 1.13 (1.09, 1.17) | 0.93 (0.92, 0.94) | 0.98 (0.96, 0.99) | 0.99 (0.99, 0.99) | 1.01 (1.01, 1.01) |
| Emergency department visit (Y vs N) | 1.02 (0.93, 1.12) | 1.34 (1.06, 1.68) | 1.12 (1.09, 1.14) | 1.31 (1.22, 1.40) | 1.15 (1.13, 1.17) | 1.12 (1.10, 1.14) |
| Hospitalization (Y vs N) | 1.14 (1.00, 1.30) | 1.19 (0.74, 1.90) | 1.31 (1.26, 1.35) | 1.31 (1.14, 1.51) | 1.16 (1.14, 1.19) | 1.16 (1.13, 1.19) |
| Institutional stay (Y vs N) | 0.30 (0.16, 0.56) | 0.49 (0.19, 1.28) | 1.15 (1.04, 1.28) | 0.88 (0.65, 1.18) | 1.15 (1.13, 1.17) | 1.19 (1.17, 1.22) |
| Comorbidity Score[d] | 0.91 (0.88, 0.94) | 0.98 (0.84, 1.13) | 0.94 (0.93, 0.96) | 1.07 (1.01, 1.13) | 0.88 (0.88, 0.89) | 0.91 (0.90, 0.91) |
| Prior hyperkalemia dx (Y vs N) | 0.19 (0.06, 0.60) | 0.05 (0.01, 0.40) | 0.73 (0.56, 0.95 | 1.80 (0.90, 3.57) | 0.74 (0.67, 0.81) | 1.24 (1.12, 1.37) |
| CKD I-IV (Y vs N) | 0.60 (0.51, 0.71) | 0.85 (0.43, 1.68) | 0.49 (0.45, 0.53) | 0.64 (0.47, 0.88) | 0.30 (0.30, 0.31) | 0.46 (0.44, 0.48) |
| MI / Stroke (Y vs N) | 1.05 (0.85, 1.30) | 1.92 (1.01, 3.64) | 1.41 (1.31, 1.51) | 3.57 (2.72, 4.68) | 0.91 (0.89, 0.94) | 0.89 (0.86, 0.92) |

| Characteristic [a] | Site 1 N=8,497 | | Site 2 N=56,266 | | Site 3 N=133,502 | |
|---|---|---|---|---|---|---|
| | Odds Ratio (95% Confidence Interval) [b] | | Odds Ratio (95% Confidence Interval) [b] | | Odds Ratio (95% Confidence Interval) [b] | |
| | Unadjusted | Adjusted [c] | Unadjusted | Adjusted [c] | Unadjusted | Adjusted [c] |
| Any Rx dispensing, increases $K^+$ [e] (Y vs N) | 0.89 (0.79, 0.99) | 1.95 (1.31, 2.92) | 0.89 (0.86, 0.93) | 1.31 (1.10, 1.56) | 0.94 (0.92, 0.96) | 1.07 (1.04, 1.09) |
| Any Rx dispensing, decreases $K^+$ [e] (Y vs N) | 0.79 (0.59, 1.05) | 1.66 (0.67, 4.13) | 0.85 (0.76, 0.95) | 0.96 (0.62, 1.47) | 0.92 (0.88, 0.97) | 1.04 (0.98, 1.09) |

[a] Determined over the 183 days prior to the cohort entry date
[b] Missing baseline glucose laboratory test results is the dependent variable
[c] Adjusted for all variables shown in table
[d] Gagne et al[69]
[e] Dispensed in 100 days prior to the cohort entry date; medication list is in Appendix D

### d. Baseline Confounder Test Case 2 analytic Approaches to handle missing serum creatinine result values using inverse probability of treatment weighting analyses: propensity scores with multiple imputation (regression method) or indicator of missing methods

We chose to initially test methods similar to those used in the baseline confounder glucose results test case as well as to add another method (inverse probability treatment weighting using PS) with analyses unique to this test case. PS models are commonly used for confounder adjustment. We explored both MI and an indicator of missing data to account for missing data in the analyses (additional details below). For all comparisons, we modeled the outcome of time to hyperkalemia diagnosis within one year in Cox Proportional Hazards models.

Missing data methods similar to the baseline confounder glucose results test case: We started with naïve models that explored model results either ignoring creatinine or using only the subset of persons with creatinine results available (i.e. complete case analyses). Complete case analyses are typically not a recommended method for dealing with missing data, but can be an important step in understanding the data and results. As noted by Sterne et al, when complete case and MI analyses give different results, one should attempt to understand why.[85] This includes thoughtful consideration of the plausibility of the MAR assumption given available variables and whether MNAR is an issue.

As in the glucose results test case, we imputed missing creatinine values using two methods in SAS® Proc MI (regression and predictive mean matching); see Section VIII.D.1.d. above for a more detailed description of these methods. The distribution of creatinine was highly skewed and a log-transformation improved the symmetry and normal approximation that is assumed in the regression imputation method. For each method, we created ten datasets with distinct imputed creatinine results. For this test case, MI were all completed within site and then combined for the outcome models. Site was included as a fixed effect in the outcome models. Cox proportional hazards models were run on each of the ten datasets and the results combined using methods of Rubin in Proc MIAnalyze.[70] The models described above comprise Models 1 through 5 in Table 13 and are summarized as:

- Model 1: Cox regression model with covariate adjustment except serum creatinine not included, full cohort (i.e. regardless of serum creatinine availability)
- Model 2: Cox regression model with covariate adjustment, serum creatinine not included but only persons with laboratory values retained (same cohort members as Model 3)
- Model 3: Cox regression model with covariate adjustment including serum creatinine (i.e. complete case analyses)

- Model 4: Cox regression model with MI regression method for missing serum creatinine values (MI completed within site; outcome model runs combined sites), Full cohort
- Model 5: Cox regression model with MI using predictive mean matching method for missing serum creatinine values (MI completed within site; outcome model runs combined sites), Full cohort

Additional missing data methods examined in this test case of baseline missing creatinine results included two methods of accounting for missing creatinine within PS models: 1) using a missing-indicator variable and 2) using multiply imputed creatinine. The outcome for the PS model was high dose lisinopril vs low dose lisinopril.

The first method using a missing-indicator variable is intuitively appealing and easy to implement. In a PS setting, the probability of treatment is modeled, and that probability might depend both on the decision to order a laboratory test, and on the test result value itself.[86] For this method, we created a dummy variable equal to 1 if creatinine was missing and equal to zero otherwise. We created a new creatinine that had the original value if it was non-missing and was set to a constant otherwise.

In addition to the missing indicator and creatinine variable described above, the PS model used the covariates in Table 11; individual comorbidities with very low frequencies (< 0.01%) were not included in the model. The outcome for the PS model was high dose lisinopril versus low dose lisinopril. The predicted probabilities of treatment were output and used to create stabilized inverse probability treatment weights that were included in the Cox Proportional Hazards model examining time to hyperkalemia diagnosis.[87]

For the second method, we used the creatinine values imputed by the regression method (as in Model 4) and included these in the PS model with an outcome of high dose versus low dose lisinopril. A PS was created for each of the ten imputation datasets. As above, the predicted probabilities were saved and used to create stabilized inverse probability treatment weights for the hyperkalemia outcome model. The outcome model was run for each of the ten datasets and the survival model results combined with Proc MIAnalyze in SAS®.

The PS modeling was completed first using main effect variables only. Because the data from all sites were combined, these initial models assume comparable associations for variables with high versus low dose lisinopril at all sites. A second set of models were run that included all site by variable interactions to create PS that were 'site-specific'. In addition, we executed models without creatinine to compare to results from the models that included creatinine using the missing data methods noted above.

These steps resulted in five models (Table 13). All five used stabilized inverse probability treatment weights but with different inputs into the PS models. These models were:

- Model 6: PS based on combined data from 3 sites; creatinine not included
- Model 7: Site specific PS; creatinine not included
- Model 8: PS based on combined data from 3 sites; creatinine included with indicator variable method for missing data
- Model 9: Site specific PS; creatinine included with indicator variable method for missing data
- Model 10: Site specific PS; MI using regression method for missing creatinine

### e. Baseline Confounder Test Case 2 results and discussion

The results from the Cox Proportional Hazards regression models are shown in Models 1 through 5 in Table 13. These models adjusted for potential confounders. In all five models, persons initiated on higher dose lisinopril were less likely to have a hyperkalemia diagnosis within the following year. The HR was closest to 1 but still significant in Model 1 (0.93 95% CI 0.88-0.99) which examined the full cohort without accounting for creatinine result value. HR and 95%CI were identical for the two MI methods; adjusting for creatinine in these models moved the HR slightly further from 1 (HR 0.88 [95% CI 0.83, 0.94] models 4 and 5). The subset with creatinine results available had HR that were further from 1 prior to including creatinine (HR 0.89, model 2); including creatinine moved estimates even further from 1 (HR 0.84 model 3).

**Table 13. Baseline Confounder Test Case 2, Hyperkalemia Diagnosis after Lisinopril Initiation: Cox Proportional Hazards Models examining the Association of High versus Low Dose Lisinopril with Hyperkalemia Diagnosis within One Year**

| | Models with Regression Covariate Adjustments | | | | | Models with Inverse Probability Treatment Weight Covariate Adjustments | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Model 1** | **Model 2** | **Model 3** | **Model 4** | **Model 5** | **Model 6** | **Model 7** | **Model 8** | **Model 9** | **Model 10** |
| Baseline serum creatinine in model? | No | No | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| Missing data method | None | None | None (complete case) | MI, regression method | MI, predictive mean matching | None | None | Indicator variable method in PS model | Indicator variable method in PS model | MI, regression method, results in PS model |
| Site specific models (for imputation or PS) | NA | NA | NA | Yes | Yes | No | Yes [a] | No | Yes [a] | Yes |
| Number in Model | 198,265 | 109,053[b] | 109,053 | 198,265 | 198,265 | 198,265 | 198,265 | 198,265 | 198,265 | 198,265 |
| HR (95% CI) lisinopril 20+mg vs < 20mg | 0.93 (0.88, 0.99) | 0.89 (0.82, 0.97) | 0.84 (0.77, 0.92) | 0.88 (0.83, 0.94) | 0.88 (0.83, 0.94) | 0.97 (0.92, 1.03) | 0.98 (0.93, 1.04) | 0.94 (0.89, 0.99) | 0.95 (0.89, 1.00) | 0.96 (0.91, 1.02) |
| HR (95% CI) Log Serum creatinine | | | 4.14[d] (3.72, 4.59) | 4.88 (4.47, 5.32) | 4.17 (3.74, 4.64) | | | | | |
| Trimmed PS[b] :N in model | | | | | | 194,013 | 194,331 | 194,376 | 193,908 | 194,499 |

| | Models with Regression Covariate Adjustments | | | | | Models with Inverse Probability Treatment Weight Covariate Adjustments | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 |
| Trimmed PS [b] HR (95% CI) lisinopril 20+mg vs < 20mg | | | | | | 0.96 (0.91, 1.02) | 0.95 (0.90, 1.01) | 0.93 (0.88, 0.99) | 0.91 (0.86, 0.97) | 0.94 (0.88, 0.99) |

[a] Models included all variable by site interactions (regardless of significance) to create propensity scores (PS) that accounted for site specific associations
[b] Selected for sample with serum creatinine results available but serum creatinine not in model
[c] PS trimmed to remove persons in lowest and highest ~1% where there was less overlap of distributions
[d] Hazard ratios appear high because Log serum creatinine has standard deviation=0.31 (i.e. <1); HR for a delta of 0.31 = 1.55 (95%CI 1.50, 1.60)

For each of these five models, we examined overlap of the PS. Overlap was generally reasonable except for some values at the tails. Trimming only a small percentage (i.e. ~< 1% and ~> 99%) removed the questionable areas. Results for both full models and trimmed models are in Table 13.

As shown in Models 6 – 10 in Table 13, we saw little difference in the PS that combined sites versus PS that were site specific (e.g. comparing Model 6 and 7 or Model 8 and 9). Despite this empiric result, allowing associations to vary by site would likely have advantages when employing PS with few downsides in settings where sample sizes are large.

All three models that included creatinine (Models 8-10) resulted in the HR moving slightly away from 1 (from 0.97-0.98 to 0.94-0.96). Similarly, all trimmed models consistently moved the HR slightly further from 1 (e.g. 0.96 to 0.94 for Model 10). The varying samples in the trimmed models make comparisons between models questionable but in general results appear similar. For example, results for trimmed Model 8 and trimmed Model 10 are nearly identical. As expected, complete case analysis (Model 3), produced an estimate quite different from other models.

In summary, we saw few differences related to the missing data methods used in this test case. Based on prior research recommendations, we consider MI models as appropriate when the MAR assumption is reasonable. We saw larger differences in results related to different analytic methods.

## 3. Baseline Confounder Test Case: Baseline INR in current warfarin users starting a potentially interacting antimicrobial medication and risk of bleeding

### a. Baseline Confounder Test Case 3 cohort development

Warfarin is an anticoagulant prescribed for prophylaxis and treatment of thromboembolic conditions. An INR is the laboratory test used to monitor the adequacy of the anticoagulant effect of warfarin. For most patients, INR monitoring is recommended to be conducted every four weeks or monthly.

This test case examined approaches that included the INR result value in patients taking warfarin who were newly-dispensed either an antibiotic that has the potential to interact with warfarin to increase the INR result value or an antibiotic that is not considered to interact with warfarin. That is, the medical product exposure of interest was initiation of selected antimicrobials in patients undergoing chronic warfarin therapy. Antimicrobials were categorized into two groups: those that interact with warfarin

(potentially interacting) and those that do not (non-interacting). Antimicrobials considered as interacting with the potential to increase bleeding risk included: fluconazole, itraconazole, ketoconazole, miconazole, ciprofloxacin, levofloxacin, moxifloxacin, norfloxacin, ofloxacin, azithromycin, erythromycin, sulfamethoxazole, sulfisoxazole, tetracycline, doxycycline, demeclocycline, chloramphenicol, isoniazid, metronidazole, and neomycin. The non-interacting comparator antimicrobials included: cephalexin, clindamycin, trimethoprim (only products not in combination with sulfamethoxazole).

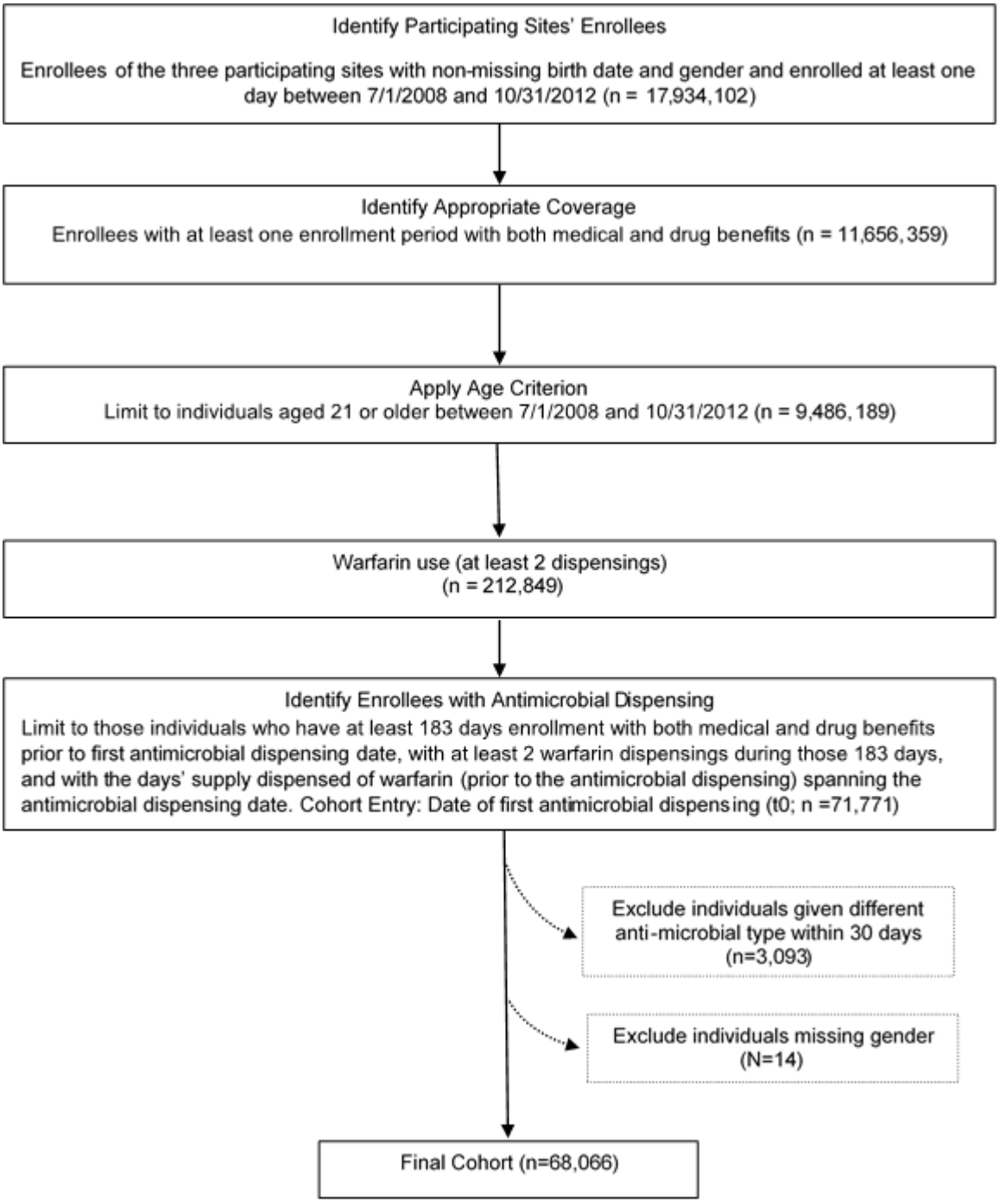
Baseline INR monitoring was defined as any INR result value up to 30 days before and including the cohort entry date. If more than one INR monitoring within days – 30 and the cohort entry date, we retained the relative date closest to the cohort entry date that INR monitoring occurred. INR results were identified and used from outpatient, emergency department, and inpatient settings. As with all test cases in this report, INR were not consistently available; the primary goal was to describe varied methods of dealing with the missing laboratory results data.

- Research Questions: In a cohort of patients taking chronic warfarin therapy,
    - Does inclusion of a baseline INR laboratory test result as a confounder in the PS model affect estimates or CI of the risk of bleeding among patients who are newly-started on an interacting or a non-interacting antimicrobial?
    - Does the choice of missing data method affect the point estimate or CI for the outcome analysis?
    - Do the results differ if the analysis is done using a pooled approach compared to stratifying by site?

Cohort members were adults $\geq$21 years of age taking chronic warfarin therapy who were members of one of the participating data partner sites between 1/1/2008 and 11/30/2013. Chronic warfarin therapy was defined as requiring at least two dispensings of warfarin prior to the dispensing date of the antimicrobial of interest (i.e., warfarin therapy started prior to the antimicrobial). The cohort entry date was the dispensing date of the antimicrobial. The days' supply dispensed of the last dispensing of warfarin prior to the cohort entry date must have spanned the cohort entry date. All cohort members were required to have medical and drug benefit coverage for at least 183 days prior to the antimicrobial dispensing. Figure 3 provides a diagram showing selection steps that resulted in the final cohort N = 68,066.

The outcome of interest was a coded bleeding/hemorrhage diagnosis from the inpatient setting within the first 30 days after the antimicrobial dispensing. A broad list of included bleeding diagnosis codes was developed from published literature (Appendix D).[88, 89] Bleeding diagnoses associated with traumatic injury were excluded.

**Figure 3. Test Case Cohort for Baseline Confounder Adjustment Test Case 3, Baseline INR in Current Warfarin Users Starting an Antimicrobial Medication [a]**




[a] Definitions and codes are in Appendix D

### b. Baseline Confounder Test Case 3 descriptive analysis of missing laboratory results

Table 14 and Table 15 show summary statistics for key variables, stratified by treatment group (Table 14) and by site (Table 15). We first describe key observations from Table 14. There were 68,066 subjects overall, with over twice as many in the interacting antimicrobial group compared with the non-interacting antimicrobial group. The outcome, any inpatient bleed within 30 days, occurs in ~ 6% of the non-interacting antimicrobial group and in ~ 7% of the interacting antimicrobial group. INR was missing at a greater rate in the interacting antimicrobial group (60% versus 47% in the non-interacting antimicrobial group). The distribution of most variables was similar between groups. There was a large difference in antimicrobial type of treatment by site – with site 3 prescribing interacting antimicrobials at a higher rate than sites 1 or 2.

**Table 14. Characteristics of Individuals in the Test Case 3 Population, Warfarin Users Starting an Antimicrobial Agent**

| Variable | Non-Interacting N=20,596 (30.3%) | Interacting N=47,470 (69.7%) | Overall N=68,066 |
|---|---|---|---|
| Bleeding diagnosis, inpatient setting within 30 days | 1251 (6.1) | 3309 (7.0) | 4560 (6.7) |
| INR laboratory test result available | 10,929 (53.1) | 19,011 (40.0) | 29,940 (44.0) |
| INR result, mean (SD) | 2.4 (0.9) | 2.4 (1.0) | 2.4 (0.9) |
| Prior history of any bleed | 4717 (22.9) | 11774 (24.8) | 16491 (24.2) |
| Prior history of inpatient bleed | 2638 (12.8) | 7373 (15.5) | 10011 (14.7) |
| Female sex | 9145 (44.4) | 22,279 (46.9) | 31,424 (46.2) |
| Age in years, mean (SD) | 72.9 (11.6) | 73.3 (10.9) | 73.1 (11.2) |
| Race | | | |
| White | 17,604 (85.5) | 39,656 (83.5) | 57,260 (84.1) |
| Black | 1112 (5.4) | 3452 (7.3) | 4564 (6.7) |
| Other | 708 (3.4) | 1469 (3.1) | 2177 (3.2) |
| Unknown | 1172 (5.7) | 2893 (6.1) | 4065 (6.0) |
| Hispanic ethnicity (Y vs N or unknown) | 838 (4.1) | 1573 (3.3) | 2411 (3.5) |
| Year of cohort entry | | | |
| 2008 | 1135 (5.5) | 2915 (6.1) | 4050 (6.0) |
| 2009 | 3474 (16.9) | 8561 (18.0) | 12,035 (17.7) |
| 2010 | 3921 (19.0) | 9181 (19.3) | 13,102 (19.2) |
| 2011 | 3957 (19.2) | 9403 (19.8) | 13,360 (19.6) |
| 2012 | 4399 (21.4) | 9590 (20.2) | 13,989 (20.6) |
| 2013 | 3710 (18.0) | 7820 (16.5) | 11,530 (16.9) |
| Site | | | |
| 1 | 1448 (7.0) | 2508 (5.3) | 3956 (5.8) |
| 2 | 7845 (38.1) | 10,845 (22.8) | 18,690 (27.5) |
| 3 | 11,303 (54.9) | 34,117 (71.9) | 45,420 (66.7) |
| INR laboratory procedure code (CPT) | 19,176 (93.1) | 42,913 (90.4) | 62,089 (91.2) |
| Additional antimicrobial dispensing within 30 days | 43 (0.2) | 1437 (3.0) | 1480 (2.2) |
| Total number of unique medication classes dispensed, mean (SD)[c] | 7.7 (4.0) | 8.1 (4.0) | 7.9 (4.0) |
| Dispensing of non-antimicrobial drug that can increase anticoagulant effect/bleeding risk of warfarin | 7965 (38.7) | 19,341 (40.7) | 27,306 (40.1) |
| Dispensing of non-antimicrobial drug that can decrease anticoagulant effect/bleeding risk of warfarin | 278 (1.3) | 714 (1.5) | 992 (1.5) |
| Number of ambulatory medical visits during baseline, mean (SD)[c] | 9.0 (7.5) | 10.0 (7.9) | 9.7 (7.8) |
| Emergency department visit during baseline, N (%) yes [c] | 19.5 | 18.0 | 18.4 |
| Hospitalization during baseline, N (%) yes [c] | 23.6 | 22.5 | 22.9 |

| Variable | Non-Interacting N=20,596 (30.3%) | Interacting N=47,470 (69.7%) | Overall N=68,066 |
|---|---|---|---|
| Institutional stay during baseline, N (%) yes [c] | 11.0 | 13.4 | 12.7 |
| Comorbidity score, mean (SD) | 1.8 (2.2) | 1.9 (2.2) | 1.9 (2.2) |
| Individual comorbidities [c, d] | | | |
| Alcohol abuse | 267 (1.3) | 553 (1.2) | 820 (1.2) |
| Anemia | 630 (3.1) | 1705 (3.6) | 2335 (3.4) |
| Arrhythmia | 11,798 (57.3) | 29,229 (61.6) | 41,027 (60.3) |
| Coagulation disorder | 431 (2.1) | 960 (2.0) | 1391 (2.0) |
| Congestive heart failure | 1302 (6.3) | 3561 (7.5) | 4863 (7.1) |
| Diabetes, complicated | < 6 (0) | 6 (0) | 7 (0.0) |
| Dementia | 226 (1.1) | 648 (1.4) | 874 (1.3) |
| HIV/AIDS | 14 (0.1) | 50 (0.1) | 64 (0.1) |
| Hypertension | 13,567 (65.9) | 32,372 (68.2) | 45,939 (67.5) |
| Hemiplegia | < 6 (0) | < 6 (0) | < 6 (0) |
| Liver disease | 308 (1.5) | 701 (1.5) | 1009 (1.5) |
| Metastatic cancer | < 6 (0) | < 6 (0) | < 6 (0) |
| Psychosis | < 6 (0) | 12 (0) | 13 (0) |
| Pulmonary disease, chronic | 2203 (10.7) | 7474 (15.7) | 9677 (14.2) |
| Pulmonary circulation disorder | < 6 (0) | 15 (0) | 20 (0) |
| Peripheral vascular disease | 2008 (9.7) | 5077 (10.7) | 7085 (10.4) |
| Renal failure | 338 (1.6) | 904 (1.9) | 1242 (1.8) |
| Tumor, any | 845 (4.1) | 1851 (3.9) | 2696 (4.0) |
| Weight loss | 31 (0.2) | 74 (0.2) | 105 (0.2) |

From Table 15 we see that there were 3956; 18,690; and 45,420 subjects in sites 1, 2, and 3, respectively. About 75% of subjects at site 3 were in the interacting group, compared with 63% in site 1 and 58% in site 2. The outcome rate was lowest in site 1 (1.6%), but fairly similar at sites 2 (6.4%) and 3 (7.3%).

The most striking difference among sites is in the rate of missing INR data. Most subjects in sites 1 and 2 have an INR value, approximately 92% in each site. At site 3, only 20% of subjects have an INR value. This is likely related to the fact that site 3 is a large national insurer site that contracts with only some of the laboratory service vendors that provide laboratory services to their enrollees. The national insurer only receives laboratory results values from some laboratory service vendors; thus, it only has laboratory test results data for only some of their patients. In contrast, sites 1 and 2 are integrated delivery systems with laboratory facilities available within their medical offices and with laboratory test results channeled into their electronic medical record databases for essentially all of their enrollees. Thus, the inclusion of INR in statistical analyses was based almost entirely on observed data in sites 1 and 2, and almost entirely on imputed data in site 3. The sites also differ on proportions of patients with several other potential confounders.

**Table 15. Characteristics of Individuals in the Baseline Confounder Test Case 3 Population, Warfarin Users Starting an Antimicrobial Agent by Data Partner Site**

| Variable | Site 1 N=3956 (5.8) | Site 2 N=18,690 (27.5) | Site 3 N=45,420 (66.7) | Overall N=68,066 |
|---|---|---|---|---|
| N (%) of individuals with interacting antimicrobial | 2508 (63.4) | 10,845 (58.0) | 34,117 (75.1) | 47,470 (69.7) |
| Bleeding diagnosis, inpatient setting within 30 days | 63 (1.6) | 1193 (6.4) | 3304 (7.3) | 4560 (6.7) |
| INR laboratory test result available | 3675 (92.9) | 17,193 (92.0) | 9072 (20.0) | 29,940 (44.0) |
| INR result, mean (SD) | 2.4 (0.8) | 2.4 (0.8) | 2.4 (1.2) | 2.4 (0.9) |
| Prior history of any bleed | 803 (20.3) | 3843 (20.6) | 11,845 (26.1) | 16,491 (24.2) |
| Female sex | 1906 (48.2) | 8698 (46.5) | 20,820 (45.8) | 31,424 (46.2) |
| Age in years, mean (SD) | 71.5 (12.9) | 72.5 (12.7) | 73.6 (10.3) | 73.1 (11.2) |
| Race | | | | |
| White | 3355 (84.8) | 15,533 (83.1) | 38,372 (84.5) | 57,260 (84.1) |
| Black | 116 (2.9) | 1000 (5.4) | 3448 (7.6) | 4564 (6.7) |
| Other | 50 (1.3) | 1598 (8.6) | 529 (1.2) | 2177 (3.2) |
| Unknown | 435 (11.0) | 559 (3.0) | 3071 (6.8) | 4065 (6.0) |
| Hispanic ethnicity (Y vs N or unknown) | 222 (5.6) | 1730 (9.3) | 459 (1.0) | 2411 (3.5) |
| Year of cohort entry | | | | |
| 2008 | 219 (5.5) | 1159 (6.2) | 2672 (5.9) | 4050 (6.0) |
| 2009 | 729 (18.4) | 3491 (18.7) | 7815 (17.2) | 12035 (17.7) |
| 2010 | 790 (20.0) | 3681 (19.7) | 8631 (19.0) | 13102 (19.2) |
| 2011 | 815 (20.6) | 3577 (19.1) | 8968 (19.7) | 13360 (19.6) |
| 2012 | 815 (20.6) | 3676 (19.7) | 9498 (20.9) | 13989 (20.6) |
| 2013 | 588 (14.9) | 3106 (16.6) | 7836 (17.3) | 11530 (16.9) |
| INR laboratory procedure code (CPT) | 3939 (99.6) | 18.416 (98.5) | 39.734 (87.5) | 62.089 (91.2) |
| Additional antimicrobial dispensing within 30 days | 91 (2.3) | 435 (2.3) | 954 (2.1) | 1480 (2.2) |
| Number of unique medication classes dispensed, mean (SD)[c] | 7.4 (4.0) | 7.8 (4.0) | 8.0 (4.0) | 7.9 (4.0) |
| Dispensing of non-antimicrobial drug that can increase anticoagulant effect/bleeding risk of warfarin | 1308 (33.1) | 6630 (35.5) | 19368 (42.6) | 27306 (40.1) |
| Dispensing of non-antimicrobial drug that can decrease anticoagulant effect/bleeding risk of warfarin | 67 (1.7) | 186 (1.0) | 739 (1.6) | 992 (1.5) |
| Number of ambulatory medical visits during baseline, mean (SD)[c] | 5.3 (4.7) | 6.2 (6.3) | 11.6 (7.9) | 9.7 (7.8) |
| Emergency department visit during baseline, N (%) yes [c] | 669 (16.9) | 5245 (28.1) | 6632 (14.6) | 12,546 (18.4) |
| Hospitalization during baseline, N (%) yes[c] | 783 (19.8) | 3562 (19.1) | 11,209 (24.7) | 15,554 (22.9) |
| Institutional stay during baseline, N (%) yes[c] | 142 (3.6) | 600 (3.2) | 7876 (17.3) | 8618 (12.7) |
| Comorbidity score, mean (SD) | 2.0 (2.3) | 1.8 (2.2) | 1.9 (2.2) | 1.9 (2.2) |
| Individual comorbidities [c,d] | | | | |
| Alcohol abuse | 65 (1.6) | 299 (1.6) | 456 (1.0) | 820 (1.2) |
| Anemia | 104 (2.6) | 565 (3.0) | 1666 (3.7) | 2335 (3.4) |
| Arrhythmia | 1783 (45.1) | 9662 (51.7) | 29,582 (65.1) | 41,027 (60.3) |
| Coagulation disorder | 106 (2.7) | 340 (1.8) | 945 (2.1) | 1391 (2.0) |
| Congestive heart failure | 204 (5.2) | 419 (2.2) | 4240 (9.3) | 4863 (7.1) |
| Diabetes, complicated | < 6 (0) | < 6 (0) | 7 (0) | 7 (0) |
| Dementia | 27 (0.7) | 157 (0.8) | 690 (1.5) | 874 (1.3) |

| Variable | Site 1 N=3956 (5.8) | Site 2 N=18,690 (27.5) | Site 3 N=45,420 (66.7) | Overall N=68,066 |
|---|---|---|---|---|
| HIV/AIDS | < 6 (0) | 31 (0.2) | 32 (0.1) | 64 (0.1) |
| Hypertension | 2016 (51.0) | 11,287 (60.4) | 32,636 (71.9) | 45,939 (67.5) |
| Hemiplegia | < 6 (0) | < 6 (0) | < 6 (0) | < 6 (0) |
| Liver disease | 82 (2.1) | 297 (1.6) | 630 (1.4) | 1009 (1.5) |
| Metastatic cancer | <6 (0) | <6 (0) | <6 (0) | <6 (0) |
| Psychosis | <6 (0) | <6 (0) | 12 (0.0) | 13 (0.0) |
| Pulmonary disease, chronic | <6 (0) | <6 (0) | 19 (0.0) | 20 (0.0) |
| Pulmonary circulation disorder | 504 (12.7) | 1606 (8.6) | 7567 (16.7) | 9677 (14.2) |
| Peripheral vascular disease | 289 (7.3) | 1525 (8.2) | 5271 (11.6) | 7085 (10.4) |
| Renal failure | 57 (1.4) | 259 (1.4) | 926 (2.0) | 1242 (1.8) |
| Tumor, any | 129 (3.3) | 773 (4.1) | 1794 (3.9) | 2696 (4.0) |
| Weight loss | 6 (0.2) | 19 (0.1) | 80 (0.2) | 105 (0.2) |

### c. Baseline Confounder Test Case 3 methods to investigate predictors of missing INR results values

To understand the relationship between baseline variables and the probability of missing INR, we completed logistic regression analyses. First, we conducted a pooled analysis using data from all sites. The outcome was present/missing baseline INR result. We fitted univariate models in addition to a multivariate model that included all of the predictors. The results are given in Table 16. As expected, site is strongest predictor of missingness, with an OR of about 0.02 comparing sites 1 and 2 to site 3 (odds of missing was about 50 times higher in site 3 compared to other sites, even after adjusting for the other variables). In the adjusted model, variables associated with significantly lower odds of missing were non-interacting antibiotic, female, older age, Hispanic ethnicity, number of prior inpatient visits, prior history of inpatient bleed, and dispensing of a non-antimicrobial drug that can increase anticoagulant effect/bleeding risk of warfarin.

**Table 16. Baseline Confounder Test Case 3, Warfarin Users Starting an Antimicrobial Agent: Odds Ratios from Logistic Regression Models of the Probability of Missing INR as a Function of Covariates [a]**

| Variable | Unadjusted OR (95%CI) | Adjusted OR (95%CI) |
|---|---|---|
| Interacting vs non-interacting reference | 1.69 (1.64, 1.75) | 1.09 (1.04, 1.15) |
| Sex, male reference | 0.93 (0.90, 0.96) | 0.91 (0.87, 0.95) |
| Age, 10 years increments | 1.01 (0.99, 1.02) | 0.91 (0.89, 0.93) |
| Race, White reference | | |
| Unknown | 1.28 (1.20, 1.37) | 0.89 (0.81, 0.99) |
| African American | 1.22 (1.15, 1.30) | 0.92 (0.85, 1.00) |
| Other | 0.26 (0.24, 0.29) | 0.94 (0.82, 1.08) |
| Hispanic ethnicity, yes reference | 0.16 (0.15, 0.18) | 0.72 (0.61, 0.83) |
| Year of cohort entry, 2008 reference | | |
| 2009 | 0.94 (0.87, 1.01) | 0.94 (0.8 , 1.04) |
| 2010 | 0.93 (0.86, 0.99) | 0.88 (0.80, 0.98) |
| 2011 | 0.98 (0.92, 1.06) | 0.92 (0.83, 1.01) |
| 2012 | 1.00 (0.93, 1.08) | 0.91 (0.82, 1.00) |
| 2013 | 1.06 (0.98, 1.14) | 1.00 (0.90, 1.10) |
| Site, Site 3 reference | | |
| Site 1 | 0.02 (0.02, 0.02) | 0.02 (0.02, 0.02) |
| Site 2 | 0.02 (0.02, 0.02) | 0.02 (0.02, 0.02) |
| INR laboratory procedure code (CPT code), Yes reference | 0.03 (0.02, 0.03) | |
| Additional antimicrobial dispensing within 30 days, Yes reference | 0.94 (0.85, 1.05) | 1.00 (0.87, 1.16) |
| Number of unique medication classes dispensed | 1.01 (1.00, 1.01) | 0.99 (0.99, 1.00) |
| Number of ambulatory medical visits during baseline | 1.08 (1.08, 1.08) | 1.02 (1.02, 1.02) |
| Emergency department visit during baseline, Yes reference | 0.64 (0.61, 0.66) | 1.10 (1.04, 1.17) |
| Hospitalization during baseline, Yes reference | 1.14 (1.10, 1.19) | 0.85 (0.81, 0.90) |
| Institutional stay during baseline, Yes reference | 3.27 (3.10, 3.45) | 1.65 (1.54, 1.77) |
| Comorbidity score | 0.96 (0.96, 0.97) | 0.90 (0.89, 0.91) |
| Prior history of any bleed, Yes reference | 1.04 (1.01, 1.08) | 0.80 (0.7 , 0.84) |
| Any dispensing of non-antimicrobial drug that can increase anticoagulant effect/bleeding risk of warfarin, Yes reference | 1.14 (1.10, 1.17) | 0.90 (0.86, 0.94) |
| Any dispensing of non-antimicrobial drug that can decrease anticoagulant effect/bleeding risk of warfarin, Yes reference | 1.25 (1.10, 1.42) | 1.03 (0.87, 1.23) |

[a] Probability modeled is INR = Missing

**Table 17. Baseline Confounder Test Case 3, Warfarin Users Starting an Antimicrobial Agent: Odds Ratios from Logistic Regression Models of the Probability of Missing INR as a Function of Covariates, Site-specific Odds Ratios [a]**

| Variable | Site | | | | | |
|---|---|---|---|---|---|---|
| | Site 1 | | Site 2 | | Site 3 | |
| | Unadjusted OR (95%CI) | Adjusted OR (95%CI) | Unadjusted OR (95%CI) | Adjusted OR (95%CI) | Unadjusted OR (95%CI) | Adjusted OR (95%CI) |
| Interacting vs non-interacting reference | 1.29 (0.99, 1.67) | 1.28 (0.98, 1.67) | 1.08 (0.97, 1.20) | 1.12 (1.01, 1.25) | 1.05 (1.00, 1.11) | 1.08 (1.02, 1.14) |
| Sex, male reference | 0.88 (0.69, 1.12) | 0.92 (0.72 , 1.18) | 0.81 (0.73 , 0.90 ) | 0.85 (0.76, 0.95 ) | 0.94 (0.90, 0.99 ) | 0.92 (0.88, 0.96) |
| Age, 10 years increments | 0.98 (0.89, 1.07) | 1.01 (0.92, 1.12) | 0.88 (0.85, 0.92 ) | 0.89 (0.85, 0.93 ) | 0.88 (0.86, 0.90 ) | 0.90 (0.88, 0.93) |
| Race, White reference | | | | | | |
| Unknown | 1.23 (0.86, 1.76) | 1.63 (1.07 , 2.48) | 1.49 (1.14, 1.95 ) | 1.40 (1.02, 1.92 ) | 0.94 (0.86, 1.03) | 0.86 (0.77, 0.97) |
| African American | 0.85 (0.39, 1.84) | 0.81 (0.37, 1.79) | 0.92 (0.72, 1.18) | 0.87 (0.68, 1.12) | 0.91 (0.84, 0.99) | 0.93 (0.85, 1.02) |
| Other | 0.27 (0.04, 1.96) | 0.26 (0.04, 1.94) | 0.95 (0.78, 1.15) | 0.87 (0.72, 1.06) | 1.01 (0.81, 1.25) | 1.01 (0.81, 1.26) |
| Hispanic ethnicity, Yes reference | 0.67 (0.36, 1.24) | 0.46 (0.23, 0.95) | 0.96 (0.80, 1.16) | 0.84 (0.68, 1.05) | 0.41 (0.34, 0.49 ) | 0.56 (0.45, 0.70) |
| Year of cohort Entry, 2008 reference | | | | | | |
| 2009 | 0.73 (0.35, 1.49) | 0.80 (0.39, 1.64) | 1.03 (0.80, 1.34) | 1.02 (0.79, 1.32) | 0.92 (0.82, 1.03) | 0.93 (0.83, 1.04) |
| 2010 | 1.17 (0.59, 2.30) | 1.31 (0.66, 2.59) | 0.99 (0.76, 1.28) | 0.95 (0.74 , 1.24) | 0.84 (0.75, 0.94) | 0.85 (0.76, 0.96) |
| 2011 | 1.72 (0.89, 3.31) | 1.93 (1.00, 3.75) | 1.11 (0.86, 1.43) | 1.07 (0.82, 1.38) | 0.85 (0.76, 0.96) | 0.86 (0.77, 0.96) |
| 2012 | 1.92 (1.00, 3.68) | 2.17 (1.12, 4.18) | 1.21 (0.94, 1.56) | 1.17 (0.91, 1.52) | 0.83 (0.74, 0.93) | 0.83 (0.74, 0.93) |
| 2013 | 1.91 (0.98, 3.73) | 2.17 (1.11, 4.26) | 1.48 (1.15, 1.91) | 1.43 (1.11, 1.85) | 0.89 (0.80, 1.00) | 0.89 (0.79, 1.00) |
| INR laboratory procedure code (CPT code), Yes reference | 0.00 (0.00, ) | NA | 0.00 (0.00, 0.00) | NA | 0.09 (0.07, 0.10) | NA |
| Additional antimicrobial dispensing within 30 days, Yes reference | 0.92 (0.40, 2.13) | 0.88 (0.37, 2.06) | 0.55 (0.35, 0.86) | 0.51 (0.32, 0.81) | 1.14 (0.96, 1.34) | 1.13 (0.96, 1.34) |
| Number of unique medication classes dispensed | 0.99 (0.96, 1.02) | 1.04 (1.00, 1.08) | 0.96 (0.94, 0.97) | 0.99 (0.97, 1.00) | 0.99 (0.99, 1.00) | 0.99 (0.99, 1.00) |
| Number of ambulatory medical visits during baseline | 0.94 (0.91, 0.97) | 0.96 (0.92, 0.99) | 0.96 (0.95, 0.97) | 0.98 (0.97, 0.99) | 1.01 (1.01, 1.02) | 1.03 (1.02, 1.03) |
| Emergency department visit during baseline, Yes reference | 0.68 (0.47, 0.98) | 0.83 (0.57, 1.22) | 0.75 (0.67, 0.85) | 0.90 (0.78, 1.04) | 1.25 (1.16, 1.33) | 1.19 (1.11, 1.28) |
| Hospitalization during baseline, Yes reference | 0.56 (0.39, 0.80) | 0.80 (0.54, 1.18) | 0.67 (0.58, 0.78) | 0.67 (0.56, 0.80) | 0.92 (0.88, 0.97) | 0.88 (0.83, 0.94) |
| Institutional stay during baseline, Yes reference | 0.18 (0.04, 0.73) | 0.27 (0.06, 1.13) | 1.29 (0.98, 1.69) | 2.26 (1.67, 3.07) | 1.52 (1.42, 1.63) | 1.60 (1.49, 1.72) |
| Comorbidity score | 0.90 (0.85, 0.95) | 0.92 (0.86, 0.98) | 0.94 (0.91, 0.96) | 0.99 (0.96, 1.02) | 0.91 (0.90, 0.92) | 0.88 (0.87, 0.89) |
| Prior history of any bleed, Yes reference | 0.47 (0.32, 0.68) | 0.50 (0.34, 0.74) | 0.62 (0.53, 0.72) | 0.70 (0.61, 0.82) | 0.78 (0.74, 0.82) | 0.82 (0.78, 0.86) |

| Variable | Site | | | | | |
|---|---|---|---|---|---|---|
| | Site 1 | | Site 2 | | Site 3 | |
| | Unadjusted OR (95%CI) | Adjusted OR (95%CI) | Unadjusted OR (95%CI) | Adjusted OR (95%CI) | Unadjusted OR (95%CI) | Adjusted OR (95%CI) |
| Any dispensing of non-antimicrobial drug that can increase anticoagulant effect/bleeding risk of warfarin, Yes reference | 0.85 (0.66, 1.11) | 0.90 (0.68, 1.20) | 0.73 (0.65, 0.82) | 0.84 (0.75, 0.95) | 0.89 (0.85, 0.93) | 0.91 (0.86, 0.95) |
| Any dispensing of non-antimicrobial drug that can decrease anticoagulant effect/bleeding risk of warfarin, Yes reference | 0.83 (0.30, 2.29) | 0.92 (0.32, 2.62) | 0.45 (0.21, 0.95 ) | 0.55 (0.25, 1.18) | 1.06 (0.88, 1.27) | 1.09 (0.90, 1.32) |
| [a] Probability modeled is INR = Missing | | | | | | |

Similar analysis was also carried out stratified by site (**Table 17**). We highlight some differences across sites. At sites 1 and 2 the odds of missing INR appear to increase over time, but the odds of missing INR were more consistent over time at site 3. At site 2 additional antimicrobial dispensing within 30 days had significantly lower risk of missing INR (OR 0.51), whereas at site 3 it was positively but non-significantly associated with missing INR (OR 1.13). Site associations with institutional stays differed across sites: OR of 0.27, 2.26, and 1.60 in sites 1-3, respectively (statistically significant at sites 2 and 3).

### d. Baseline Confounder Test Case 3 – Analytic approaches to handle missing INR result values: propensity score matching

The general approach to the strategy for assessing missingness in this test case was to employ PS matching to control for confounding. Risk differences and CI were determined from matched data.

In previous test cases to control for confounding we used regression adjustment (test case 1) and inverse probability of treatment weighting (test case 2). In this test case we use PS matching. We fit a PS model using logistic regression, with interacting versus non-interacting antimicrobial as the outcome and the variables listed in Table 14 as predictors. After we estimated a PS for each subject, we matched one-to- one using the SAS macro GMATCH. We matched in the non-interacting antimicrobial group because there were fewer subjects in that group. For each subject in the non-interacting group, we attempted to find a match from the interacting group. A match was accepted as long as it was within the caliper set to 0.2 times the standard deviation of logit of the PS as recommended in Austin.[90] To assess performance on the matching algorithm, we calculated standardized differences between the treatment groups on the matched data for the variables that were in the PS. Standardized differences are commonly used to assess balance on covariates, because unlike p-values, they should not decrease with sample size. A common rule-of-thumb is that reasonable balance is achieved if the standardized differences are less than 0.1 - 0.2 in absolute value.

Our interest was in the risk difference in the outcome between groups, after controlling for confounding via matching. We used PROC GENMOD in SAS with bleeding as the outcome, treatment group as the predictor, identity linking to obtain a risk difference, and invoked the REPEATED option to account for matching (GEE).

Because INR was a potential confounder, it was accounted for in the PS. We implemented three approaches: (1) do not include INR in the PS; (2) use the indicator method; (3) use MI (predictive mean matching). The indicator method (described previously in baseline confounder test case 2) involves creating two INR variables: an indicator that INR is observed (yes/no), and INR with missing result values set to the same value for all subjects for whom the INR result value is missing (the specific value does

not matter; we set it to 0). The same predictive mean matching approach was used as previously described in test cases 1 and 2. Ten imputed data sets were created, and for each one the PS was fitted, matching was carried out, and the outcome was analyzed. Estimates and standard errors were then combined using the methods of Rubin[70] to get a single estimate.

All analyses were conducted overall and separately by site.

Because we used a caliper, it was likely there would be subjects in the non-interacting group who did not match to a subject in the interacting group. In addition, the number of subjects we could find a match for could vary, depending on the missing data method used. For example, it is possible that subjects that appear to be a good match if we ignore INR might not be a good match if we include INR. We report sample sizes from each method in Table 18. Nearly all the 20,596 subjects in the non-interacting group were matched to a subject in the interacting group, regardless of missing data method. For example, the indicator method matched 20,579 of the 20,596 subjects (no matches for 17).

**Table 18. Baseline Confounder Test Case 3, Warfarin Users Starting an Antimicrobial Agent: Sample Sizes for Each Treatment Group Overall and by Data Partner Site, Original Sample and for Each Imputation Method**

| | Overall | | Site 1 | | Site 2 | | Site 3 | |
|---|---|---|---|---|---|---|---|---|
| | Interacting | Non-Interacting | Interacting | Non-Interacting | Interacting | Non-Interacting | Interacting | Non-Interacting |
| Original | 47,470 | 20,596 | 2,508 | 1,448 | 10,845 | 7,845 | 34,117 | 11,303 |
| **After matching** | | | | | | | | |
| Exclude INR | 20,578 | 20,578 | 1,437 | 1,437 | 7,781 | 7,781 | 11,302 | 11,302 |
| Indicator | 20,579 | 20,579 | 1,434 | 1,434 | 7,783 | 7,783 | 11,302 | 11,302 |
| Multiple Imputation (min, max) | (20,577, 20,580) | (20,577, 20,580) | (1,434, 1,436) | (1,434, 1,436) | (7,780, 7,783) | (7,780, 7,783) | (11,298, 11,302) | (11,298, 11,302) |

To assess the covariate balance after matching, we report standardized differences (Appendix E, Supplementary Tables 1 [ignoring INR], 2 [indicator method], and 3 [MI]). For every variable, both by site and overall, standardized differences were less than 0.1. Thus, matching achieved acceptable balance.

### e. Baseline Confounder Test Case 3 – Results and Discussion

Outcome analysis results are in **Table 19**. For all three missing data methods, the risk difference was positive and either statistically significant or nearly significant in pooled analysis. The point estimate ranged from 0.05, for the method that ignored INR, to 0.09, for the MI method. For example, a risk difference of 0.09 implies that an additional 9 of every 100 subjects would experience bleeding within 30 days in the interacting group compared to the non-interacting group. The 95%CI was widest for the MI method.

**Table 19. Baseline Confounder Test Case 3, Warfarin Users Starting an Interacting versus Non-Interacting Antimicrobial Agent: Differences in Risk of Any Inpatient Bleeding Outcome Diagnosis within 30 Days between Interacting and Non-Interacting Antimicrobials Prescribed to Patient Taking Chronic Warfarin Therapy [a]**

| | All Sites Combined | SITE1 | SITE2 | SITE3 |
|---|---|---|---|---|
| | SD (95%CI) | SD (95%CI) | SD (95%CI) | SD (95%CI) |
| Ignore INR | 0.05 (-0.03, 0.13) | 0.14 (-0.45, 0.74) | 0.17 (0.04, 0.30) | 0.02 (-0.08, 0.12) |
| INR indicator method | 0.07 (-0.01, 0.15) | 0.38 (-0.19, 0.95) | 0.15 (0.01, 0.28) | -0.01 (-0.12, 0.09) |
| Imputed INR | 0.09 (0.00, 0.18) | 0.30 (-0.30, 0.90) | 0.17 (0.03, 0.30) | 0.01 (-0.11, 0.13) |

[a] Risk differences and 95% CIs. Non-interacting is the reference; a positive risk difference indicates the interacting antimicrobial is associated with a higher risk of bleeding

Site-specific results are also shown in Table 19. Risk differences varied greatly by site. Sites 1 and 2 show greatly elevated bleeding risk within 30 days when an interacting vs. non-interacting antimicrobial was dispensed (statistically significant in site 2), whereas for site 1 the risk difference shows little difference in bleeding risk between interacting and non-interacting antimicrobials. The indicator method and predictive mean matching yielded very similar estimates and CI. Predictive mean matching provided slightly wider CI.

In summary, in this test case there were differences between the method that ignored missing laboratory result values and the methods that accounted for missing result values. However, the two methods that accounted for missing data provided similar results. As in previous test cases, we saw large site differences in amount of missing data, predictors of missingness, and outcome.

## 4. Additional consideration: number of datasets to impute

One decision when using MI is the number of imputed datasets to create and analyze. In these test cases, we used ten imputed datasets in the primary analyses. Here we briefly discuss considerations about the choice of number of datasets as well as sub-analyses that varied the number of replicate datasets in baseline confounder test case 2.

Enders provides a summary of recommendations related to number of datasets.[2] Although three to five have been historically recommended, Enders notes that the decrease in standard errors as the number of imputations increase may be one reason to increase the number of imputed datasets. The relative efficiency estimate provided by most MI software quantifies the potential standard error size change relative to its theoretical minimum (i.e., estimate that would occur with an infinite number of datasets). For example, for a dataset with ten imputations, a relative efficiency of 92% suggests the sampling variance (squared standard error) with infinite datasets would be 92% as large as the one estimated with ten replicates. Relative efficiency improves with the number of imputations and decreases with higher proportions of missing data. From a table of relative efficiency estimates for missing data fractions ranging from 10% to 70% and with 3, 5, 10 or 20 imputed datasets, Enders concluded the largest gains in efficiency occurred between 3 and 10 datasets. In baseline confounder test case 2, in addition to the ten imputed datasets of the primary analysis, we analyzed with 5 and with 20 imputed datasets. The results from these additional imputations are in Appendix F.

## 5. Additional consideration: technical performance/SAS runtime

We employed SAS® version 9.3 or 9.4 for all analyses for all baseline confounder test cases. The predictive mean matching method and the regression method of imputation had large differences in run times. The predictive mean matching methods took longer (several more hours) to complete than did

the regression method. As an example, both these methods were used to impute missing serum creatinine values in baseline confounder adjustment test case 2. This cohort had 198,265 members with 45% missing creatinine (n=89,212). MI runs using the regression method in SAS® version 9.4 required less than five minutes of run time and less than one minute of CPU time. By contrast, MI using predictive mean matching took over five hours and 23 minutes of CPU time. Disparities in run times are important considerations. We do not know whether such dramatic differences in run time would occur with other software packages.

## 6. Summary of the performance of the tested methods across Baseline Confounder Test Cases

Observations about the performance of the tested methods across these three baseline confounder test cases include:

- Covariate associations with missing data varied across sites. Missing data methods need to allow for this variability by, for example, conducting imputations within sites or by employing models that include site by variable interaction terms.
- Different missing data methods often provided comparable "answers."
    - Comparable results were particularly evident in comparisons of two MI methods. Missing data estimated by either predictive mean matching or regression imputation resulted in identical or very similar point and CI estimates (i.e., differing by no more than 0.01 in both test cases 1 and 2).
    - Similar results were evident for the indicator method and MI methods of accounting for missing data in the inverse PS weighted models of test case 2.
- Wider CIs were observed in the meta-analyses that combined site specific imputed results compared to pooled data analyses (test case 1). The differences in CIs were likely influenced by varied associations in the site specific imputations models. Prior literature has shown much less difference between pooled and site specific meta-analyses when data are comparably specified.[91] As noted above, missing data varied by site. We recommend site specific or other methods to appropriately account for such differences.
- In these test cases, differing analytic models impacted results more strongly than did differing missing data methods. For example, in test case 2, MI was used in both regression adjusted models and models adjusted by inverse PS weights, and parameter estimates diverged (0.88 vs 0.96; Models 4 & 10 in Table 13). In contrast, using differing missing data methods (indicator and MI) but the same analytic method of inverse PS adjustments produced comparable estimates (0.95 vs 0.96; Models 9 & 10 in Table 13).
- Laboratory test results were strongly associated with outcomes but had limited impact as confounder adjustments. Given the variability in missingness patterns for specific laboratory test results and across sites, we cannot generalize our finding that laboratory test result adjustments will not have substantial impacts on other cohorts or other analyses.
- Data partner site-specific imputations and data adjustments are recommended due to substantive differences in the missing data by site. In addition, site specific models also have the advantage of fitting the preferred distributed data analyses of Mini-Sentinel.

## E. COHORT IDENTIFICATION TEST CASES

The purpose of the Cohort Identification Test Cases was to determine whether additional individuals with a given medical condition were identified, and if so, the magnitude of the gain, when available laboratory test results were considered as criteria for cohort inclusion in addition to coded diagnoses and procedures. This is different from the purpose of the Baseline Confounder Test Cases, in which we used laboratory test results to evaluate baseline confounding and to evaluate the performance of missing data techniques. In the Cohort Identification Test Cases, the workgroup also describe the characteristics of individuals identified using differing cohort inclusion criteria, the proportions of the cohorts with laboratory results data available, and examined cohort-specific research questions.

### 1. Cohort Identification Test Case 1: enhanced identification of a pregnancy cohort through use of positive qualitative or high quantitative pregnancy test results

#### a. Cohort Identification Test Case 1 cohort development

The workgroup examined whether incorporating "pregnancy hormone" laboratory test results, including quantitative or qualitative human chorionic gonadotropin (HCG) values, increased the size of a cohort of women identified as pregnant compared to only using pregnancy-related diagnosis and procedure codes, which is the typical approach for administrative data. For women who delivered a live born infant, the workgroup also determined whether the estimated gestational length changed when laboratory test results were included.

The cohort entry date was the earliest positive indicator of pregnancy, including a positive qualitative or quantitative HCG laboratory test result, a prenatal care visit or procedure, or a coded pregnancy diagnosis or prenatal procedure. Because the purpose was to enhance identification of a cohort of pregnant women, no medical product exposure was required. The list of codes used to identify pregnancy is in Appendix G. HCG was defined as positive when the qualitative HCG result was positive or the quantitative HCG result was > 25 mIU/mL.[92] For each woman, only the first observed pregnancy episode within the study timeframe was included.

The end of pregnancy indicators of interest included term and preterm delivery of a live born infant, multiple live births, pregnancy loss including fetal death/stillbirth, ectopic and other extra-uterine pregnancies, miscarriage and therapeutic/elective abortion, disenrollment from the health plan, death, end of study timeframe, and no end of pregnancy indicator found in the dataset.

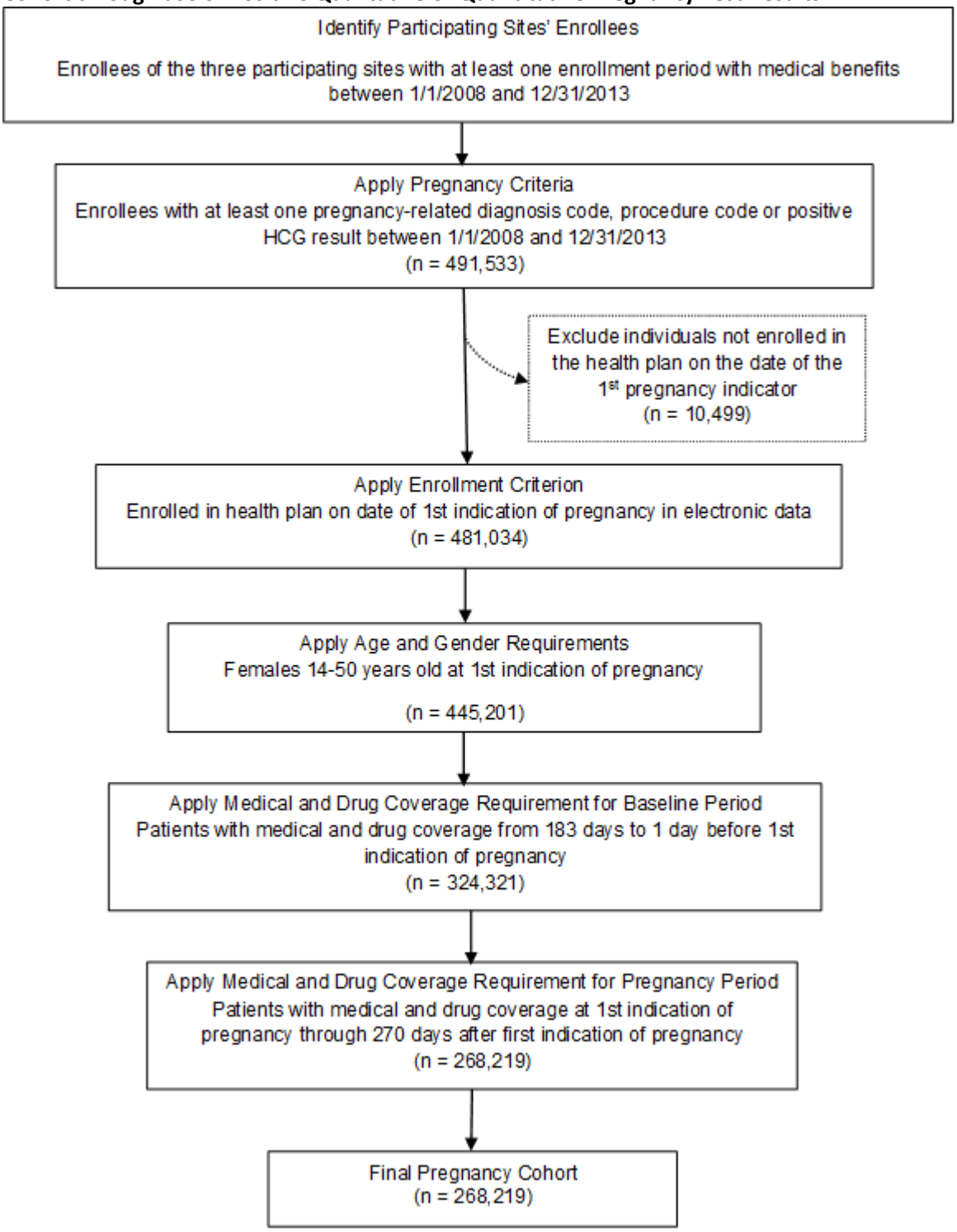
The following research questions were addressed:

1. How many pregnancies are gained by including pregnancies identified using laboratory test results that can include women with pregnancy loss or no live born delivery?
    - Determine the numbers and proportions of pregnancies detected by
        a. Diagnosis and/or procedure codes only
        b. Pregnancy laboratory test results only
        c. Both pregnancy laboratory test results and diagnosis and/or procedure codes.
    - For women with a first pregnancy that includes a pregnancy laboratory test result, describe the number of qualitative and/or quantitative HCG laboratory test results per woman.
    - Determine the numbers and proportions of pregnancies detected by each of the methods in each of the end of pregnancy categories of interest.

2. For women with a live born delivery identified in the dataset whose pregnancies had both coded diagnoses or procedures and pregnancy laboratory test results, how does the first positive qualitative or quantitative laboratory test result change the timing of when the pregnancy was first identified in the electronic data? What is the difference in how early a pregnancy is identified when pregnancy laboratory test results are incorporated?
    - What proportion of women have the pregnancy identified earlier using the laboratory test result? Summarize the lead time in days or weeks gained by considering the pregnancy laboratory test result.
    - What proportion of women have a pregnancy identified earlier using the diagnosis or procedure code? Summarize the lead time gained by having the diagnosis or procedure code.
    - Does this differ by data partner site?
    - Using the 270 day metric that has been applied in previous observation studies based only on delivery codes (delivery code date minus 270 days is considered the estimated length of gestation)[93-95]
        a. Among the women who have a pregnancy first identified from a pregnancy laboratory test result (with or without a diagnosis or procedure code), for what proportion does the date of the first positive pregnancy laboratory test result fall within the date range of the 270 day metric?
        b. Use the 270 day metric to identify the first trimester as the initial 90 days among women who have a pregnancy first identified using a pregnancy laboratory test result (with or without a diagnosis or procedure code; the laboratory test result can be prior to or after the code). Describe how many women have a laboratory test result only in the first trimester period versus diagnosis or procedure codes versus both laboratory test result and diagnosis or procedure codes.

Cohort members were women from the three participating data partner sites ages 14 through 50 years on the cohort entry date (date of first pregnancy indicator). The date range for the cohort included January 1, 2008 – December 31, 2013. Women were required to be enrolled in the health plan and to have medical and drug coverage for at least 183 days before the cohort entry date (bridging gaps up to 45 days) and for at least 270 days after the cohort entry date. Figure 4 shows selection steps that resulted in the final cohort (n = 268,219). Further information on the specifications of the test case, such as covariate data collected, is in Appendix D.

**Figure 4. Test Case for Cohort Identification Test Case 1, Enhanced Identification of a Pregnancy Cohort through use of Positive Qualitative or Quantitative Pregnancy Test Results**



### b. Cohort Identification Test Case 1 descriptive analysis of missing HCG laboratory test results

Detailed descriptions of the cohort overall, by site, and by end of pregnancy indicator are in Table 20 and Table 21. Overall, among the 268,219 women in the cohort, the mean age was 30.5 (SD 7.1) years, likely reflecting that, due to the approach to cohort construction (i.e., taking the first observed pregnancy during the study period in the dataset for each woman during the date range), the cohort included a mixture of women with first and subsequent pregnancies. As shown in Table 21, women whose pregnancies ended with abortion were younger (mean age 27.3 years) and women with extra-uterine pregnancies were older (mean age 34.0 years) than the overall cohort population age.

**Table 20. Characteristics of the Pregnancy Cohort Identification Test Case Population Overall and by Site [a]**

| Characteristic | Total, N (%) | Data Partner Site, N (%) | | |
|---|---|---|---|---|
| | | Site 1 | Site 2 | Site 3 |
| Number of women | 268,219 | 24,649 | 187,471 | 56,099 |
| End of pregnancy indicator [b] | | | | |
| Single live birth: Term | 157,076 (58.6) | 15,060 (61.1) | 107,675 (57.4) | 34,341 (61.2) |
| Single live birth: Preterm | 10,261 (3.8) | 974 (4.0) | 6,796 (3.6) | 2,491 (4.4) |
| Abortion | 21,357 (8.0) | 1,127 (4.6) | 19,991 (10.7) | 239 (0.4) |
| Miscarriage | 29,899 (11.1) | 2,637 (10.7) | 20,696 (11.0) | 6,566 (11.7) |
| Fetal death | 2,006 (0.7) | 214 (0.9) | 1,440 (0.8) | 352 (0.6) |
| Extra-uterine | 3,146 (1.2) | 240 (1.0) | 2,127 (1.1) | 779 (1.4) |
| Abnormal conception product [c] | 2,083 (0.8) | 195 (0.8) | 1,252 (0.7) | 636 (1.1) |
| Multiple births | 4,027 (1.5) | 386 (1.6) | 2,614 (1.4) | 1,027 (1.8) |
| End of pregnancy indicator not found | 38,364 (14.3) | 3,816 (15.5) | 24,880 (13.3) | 9,668 (17.2) |
| Reason end of pregnancy indicator not found | | | | |
| Death | 7 (0.0) | 0 (0.0) | 4 (0.0) | 3 (0.0) |
| End of study (estimated due date beyond 12/31/2013) | 14,287 (5.3) | 1,603 (6.5) | 10,015 (5.3) | 2,669 (4.8) |
| Unable to determine from database | 24,070 (9.0) | 2,213 (9.0) | 14,861 (7.9) | 6,996 (12.5) |
| Age in years, mean (SD) | 30.5 (7.1) | 29.7 (6.7) | 30.7 (7.3) | 30.3 (6.5) |
| Race | | | | |
| White | 103,204 (38.5) | 15,215 (61.7) | 86,856 (46.3) | 1,133 (2.0) |
| Black | 20,122 (7.5) | 1,240 (5.0) | 18,521 (9.9) | 361 (0.6) |
| Other | 49,285 (18.4) | 1,389 (5.6) | 47,863 (25.5) | 33 (0.1) |
| Unknown | 95,608 (35.6) | 6,805 (27.6) | 34,231 (18.3) | 54,572 (97.3) |
| Hispanic ethnicity (Y vs N/unknown) | 51,794 (19.3) | 4,240 (17.2) | 47,445 (25.3) | 109 (0.2) |
| Qualitative or Quantitative HCG lab result value [d] | | | | |
| No HCG lab result | 127,613 (47.6) | 7,919 (32.1) | 76,108 (40.6) | 43,586 (77.7) |
| Only negative HCG lab result | 21,368 (8.0) | 1,518 (6.2) | 18,538 (9.9) | 1,312 (2.3) |
| 1+ Positive HCG lab result; women with lab result and no pregnancy diagnosis | 13,665 (5.1) | 679 (2.8) | 12,582 (6.7) | 404 (0.7) |
| 1+ Positive HCG lab result; women with diagnosis and lab result | 105,573 (39.4) | 14,533 (59.0) | 80,243 (42.8) | 10,797 (19.2) |
| Qualitative HCG lab | | | | |
| No Qualitative HCG lab result | 175,406 (65.4) | 15,535 (63.0) | 105,342 (56.2) | 54,529 (97.2) |
| Only negative Qualitative HCG lab result | 22,882 (8.5) | 2,884 (11.7) | 19,312 (10.3) | 686 (1.2) |
| 1+ Positive Qualitative HCG lab; women with lab result and no pregnancy diagnosis | 10,213 (3.8) | 420 (1.7) | 9,692 (5.2) | 101 (0.2) |
| 1+ Positive Qualitative HCG lab; women with pregnancy diagnosis and lab result | 59,718 (22.3) | 5,810 (23.6) | 53,125 (28.3) | 783 (1.4) |

| Characteristic | Total, N (%) | Data Partner Site, N (%) | | |
|---|---|---|---|---|
| | | Site 1 | Site 2 | Site 3 |
| Quantitative HCG lab [d] | | | | |
| No Quantitative HCG lab result | 190,236 (70.9) | 11,091 (45.0) | 134,599 (71.8) | 44,546 (79.4) |
| Only negative Quantitative HCG lab result | 9,443 (3.5) | 412 (1.7) | 8,153 (4.3) | 878 (1.6) |
| 1+ Positive Quantitative HCG lab; women with lab results and no pregnancy diagnosis | 4,604 (1.7) | 361 (1.5) | 3,921 (2.1) | 322 (0.6) |
| 1+ Positive Quantitative HCG lab result; women with pregnancy diagnosis and lab result | 63,936 (23.8) | 12,785 (51.9) | 40,798 (21.8) | 10,353 (18.5) |
| Medical Utilization [e] | | | | |
| No encounters prior to first pregnancy indicator | 55 (0.0) | 2 (0.0) | 33 (0.0) | 20 (0.0) |
| Ambulatory visits, mean (SD) | 6.4 (7.1) | 6.6 (7.5) | 6.6 (7.2) | 5.7 (6.5) |
| Emergency department visits, % yes | 31,408 (11.7) | 2,534 (10.3) | 22,541 (12.0) | 6,333 (11.3) |
| Hospitalization, % yes | 10,398 (3.9) | 668 (2.7) | 8,373 (4.5) | 1,357 (2.4) |
| Institutional stay, % yes | 1,399 (0.5) | 2 (0.0) | 57 (0.0) | 1,340 (2.4) |
| HCG injections | 6,037 (2.3) | 570 (2.3) | 5,288 (2.8) | 179 (0.3) |
| Tumor that may influence HCG | 1,769 (0.7) | 77 (0.3) | 1,413 (0.8) | 279 (0.5) |

[a] Women included in the cohort have enrollment in the health plan for at least 183 days prior to the cohort entry date and 270 days after the cohort entry date. Cohort entry date is the date of first indicator of pregnancy (i.e., diagnosis code, procedure code, or positive laboratory test result)
[b] Within 270 days after the cohort entry date
[c] ICD9 codes 630, 631 (i.e. Hydatidiform mole or other abnormal products of conception)
[d] Quantitative lab considered positive if >25 mIU/mL
[e] During 183 days prior to cohort entry (earliest pregnancy indicator). Women with no HCG result or with only negative HCG result entered the cohort through a pregnancy-related diagnosis or procedure code

As shown in Table 20, 52.4% of the women had at least one HCG laboratory test result (i.e., 47.6% had no HCG laboratory test result). Importantly, this proportion included 5.1% with a positive HCG result value and no pregnancy-related diagnosis or procedure code. Availability of HCG test results varied by site, with 22.3% at site 3 having at least one HCG test result (i.e., 77.7% with no HCG lab result), to 59.4% at site 2 having at least one HCG test result, to 67.9% at site 1 having at least on HCG test result (Table 20). The 47.6% with no HCG laboratory test result included 8.9% (n = 23,741) of women with no indicator of pregnancy until the end of pregnancy indicator was coded.

Across the cohort, 62.4% had a pregnancy that ended with a full term or preterm single live birth, 1.5% ended with multiple births, 19.1% ended with miscarriage or abortion, and 0.7% ended with fetal demise (Table 20). Additionally, 1.2% of pregnancies were extra-uterine and 0.8% contained an abnormal product of conception. No end of pregnancy indicator was found for 14.3%, many of whom (5.3%) had an estimated due date beyond the end of the study period.

End of pregnancy percentages varied by site, particularly for abortion (from 0.4% at site 3 to 10.7% at site 2)(Table 20). Personal communication with site 3 suggests site 3 may less often provide insurance coverage for abortion (than do sites 1 or 2). It is also feasible there is less interest or willingness of providers affiliated with site 3 to refer members for abortions, as well as less access by members to that service. These factors may correlate with site 3 having a higher proportion of membership in conservative geographic regions of the United States that have more stringent state regulation of providers (than do sites 1 or 2).

Table 21 shows that, among the 5.1% (13,665 of 268,219) of women who had a positive HCG laboratory test result and no pregnancy-related diagnosis or procedure code, approximately 75% had a pregnancy

loss recorded (e.g., 6,953 had an abortion, 3,014 had a miscarriage, 274 had an extra-uterine pregnancy) and 22.5% (n = 3,081) had no end of pregnancy indicator recorded in the database.

In addition, relative to the overall proportion of women in the cohort (52.4% with a HCG result), a higher proportion of women with pregnancy losses had HCG results, including women with an abortion (72.8%), miscarriage (67.4%), fetal death (58.7%), extra-uterine pregnancies (65.1%), and an abnormal product of conception (69.9%). A lower proportion of women with single live births had HCG results, including 46.7% of women with term births and 48.5% of women with preterm births. Also, women with certain subsets of pregnancy losses (miscarriage, fetal death, extra-uterine pregnancy, abnormal products of conception) had more HCG laboratory test results than did women with single live births, whether or not the women also had a pregnancy diagnosis code. The highest mean (SD) number of HCG results was 4.5 (3.0) in women with extra-uterine pregnancies. Women with multiple births and those with no end of pregnancy indicator found also had a higher number of HCG results than women with single live births. In all subsets with more HCG laboratory result values available (i.e., a higher number of HCG results), the increase was driven by the number of quantitative HCG results.

**Table 21. Characteristics of the Pregnancy Cohort Identification Population by End of Pregnancy Indicator [a]**

| Characteristic | Total N = 268,219 (100%) | Single Live Birth | | Pregnancy Loss | | | | | Multiple births N = 4,027 (1.5%) | No end of pregnancy indicator found N = 38,364 (14.3%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Term N = 157,076 (58.6%) | Preterm N = 10,261 (3.8%) | Abortion N = 21,357 (8.0%) | Miscarriage N = 29,899 (11.1%) | Fetal death N = 2,006 (0.7%) | Extra-uterine N = 3,146 (1.2%) | Abnormal conception [b] N = 2,083 (0.8%) | | |
| Age in years, mean (SD) | 30.5 (7.1) | 30.5 (6.7) | 29.8 (6.2) | 27.3 (7.8) | 31.6 (6.8) | 31.4 (6.4) | 34.0 (7.8) | 31.6 (7.1) | 31.9 (5.7) | 31.1 (8.0) |
| Qualitative (QL) or Quantitative (QN) HCG lab result [c] | | | | | | | | | | |
| No HCG lab result | 127,613 (47.6) | 83,690 (53.3) | 5,282 (51.5) | 5,803 (27.2) | 9,737 (32.6) | 829 (41.3) | 1,098 (34.9) | 628 (30.1) | 1,966 (48.8) | 18,580 (48.4) |
| Only negative HCG lab result | 21,368 (8.0) | 12,741 (8.1) | 358 (3.5) | 743 (3.5) | 1,742 (5.8) | 74 (3.7) | 857 (27.2) | 75 (3.6) | 107 (2.7) | 4,671 (12.2) |
| ≥ 1 Positive HCG lab; women with lab result and no pregnancy (PG) diagnosis | 13,665 (5.1) | 28 (0.0) | 107 (1.0) | 6,953 (32.6) | 3,014 (10.1) | 16 (0.8) | 274 (8.7) | 192 (9.2) | 0 (0.0) | 3,081 (8.0) |
| N of labs, mean (SD) | 1.3 (0.8) | 1.2 (0.4) | 1.1 (0.4) | 1.1 (0.3) | 1.5 (0.9) | 1.6 (1.1) | 1.9 (1.4) | 1.6 (1.0) | 0 | 1.5 (1.3) |
| ≥ 1 Positive HCG lab; women with PG diagnosis and lab result | 105,573 (39.4) | 60,617 (38.6) | 4,514 (44.0) | 7,858 (36.8) | 15,406 (51.5) | 1,087 (54.2) | 917 (29.1) | 1,188 (57.0) | 1,954 (48.5) | 12,032 (31.4) |

| Characteristic | Total N = 268,219 (100%) | Single Live Birth | | Pregnancy Loss | | | | | Multiple births N = 4,027 (1.5%) | No end of pregnancy indicator found N = 38,364 (14.3%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Term N = 157,076 (58.6%) | Preterm N = 10,261 (3.8%) | Abortion N = 21,357 (8.0%) | Miscarriage N = 29,899 (11.1%) | Fetal death N = 2,006 (0.7%) | Extra-uterine N = 3,146 (1.2%) | Abnormal conception [b] N = 2,083 (0.8%) | | |
| N of labs, mean (SD) | 1.8 (1.3) | 1.5 (0.9) | 1.7 (1.1) | 1.5 (1.1) | 2.4 (1.5) | 2.1 (1.8) | 4.5 (3.0) | 3.0 (2.0) | 2.0 (1.2) | 2.2 (1.9) |
| QL HCG lab result | | | | | | | | | | |
| No QL HCG lab | 175,406 (65.4) | 107,961 (68.7) | 7,056 (68.8) | 7,664 (35.9) | 19,066 (63.8) | 1,265 (63.1) | 1,944 (61.8) | 1,393 (66.9) | 3,073 (76.3) | 25,984 (67.7) |
| Only negative QL HCG lab | 22,882 (8.5) | 13,873 (8.8) | 577 (5.6) | 987 (4.6) | 2,129 (7.1) | 135 (6.7) | 765 (24.3) | 168 (8.1) | 289 (7.2) | 3,959 (10.3) |
| ≥ 1 Positive QL HCG lab; women with lab result only | 10,213 (3.8) | 11 (0.0) | 26 (0.3) | 6,554 (30.7) | 1,283 (4.3) | 12 (0.6) | 113 (3.6) | 51 (2.4) | 0 (0.0) | 2,163 (5.6) |
| N of labs, mean (SD) | 1.0 (0.2) | 1.0 (0.0) | 1.0 (0.2) | 1.0 (0.1) | 1.0 (0.2) | 1.1 (0.3) | 1.1 (0.4) | 1.0 (0.2) | . (.) | 1.0 (0.2) |
| ≥ 1 Positive QL HCG lab; women with diagnosis and lab result | 59,718 (22.3) | 35,231 (22.4) | 2,602 (25.4) | 6,152 (28.8) | 7,421 (24.8) | 594 (29.6) | 324 (10.3) | 471 (22.6) | 665 (16.5) | 6,258 (16.3) |
| N of labs, mean (SD) | 1.0 (0.2) | 1.0 (0.2) | 1.1 (0.2) | 1.1 (0.2) | 1.1 (0.2) | 1.1 (0.3) | 1.1 (0.4) | 1.1 (0.3) | 1.0 (0.3) | 1.1 (0.2) |
| QN HCG lab result [c] | | | | | | | | | | |
| No QN HCG lab | 190,236 (70.9) | 119,644 (76.2) | 7,341 (71.5) | 17,313 (81.1) | 13,623 (45.6) | 1,253 (62.5) | 1,680 (53.4) | 738 (35.4) | 2,467 (61.3) | 26,177 (68.2) |
| Only negative QN HCG lab | 9,443 (3.5) | 3,775 (2.4) | 153 (1.5) | 452 (2.1) | 1,533 (5.1) | 23 (1.1) | 306 (9.7) | 61 (2.9) | 57 (1.4) | 3,083 (8.0) |
| ≥ 1 Positive QN HCG; women with lab result and no PG diagnosis | 4,604 (1.7) | 19 (0.0) | 88 (0.9) | 590 (2.8) | 2,307 (7.7) | 6 (0.3) | 263 (8.4) | 179 (8.6) | 0 (0.0) | 1,152 (3.0) |
| N of labs, mean (SD) | 1.5 (1.2) | 1.2 (0.4) | 1.0 (0.2) | 1.3 (0.8) | 1.3 (0.8) | 2.2 (1.6) | 1.5 (1.2) | 1.4 (0.9) | . (.) | 2.1 (1.7) |

| Characteristic | Total N = 268,219 (100%) | Single Live Birth | | Pregnancy Loss | | | | | Multiple births N = 4,027 (1.5%) | No end of pregnancy indicator found N = 38,364 (14.3%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Term N = 157,076 (58.6%) | Preterm N = 10,261 (3.8%) | Abortion N = 21,357 (8.0%) | Miscarriage N = 29,899 (11.1%) | Fetal death N = 2,006 (0.7%) | Extra-uterine N = 3,146 (1.2%) | Abnormal conception [b] N = 2,083 (0.8%) | | |
| ≥ 1 Positive QN HCG; women with PG diagnosis and lab results | 63,936 (23.8) | 33,638 (21.4) | 2,679 (26.1) | 3,002 (14.1) | 12,436 (41.6) | 724 (36.1) | 897 (28.5) | 1,105 (53.0) | 1,503 (37.3) | 7,952 (20.7) |
| N of labs, mean (SD) | 2.0 (1.4) | 1.7 (1.0) | 1.9 (1.1) | 1.9 (1.4) | 2.4 (1.4) | 2.3 (1.9) | 4.2 (2.9) | 2.8 (1.9) | 2.2 (1.1) | 2.4 (2.1) |
| HCG injections | 6,037 (2.3) | 2,178 (1.4) | 183 (1.8) | 82 (0.4) | 1,256 (4.2) | 81 (4.0) | 130 (4.1) | 102 (4.9) | 505 (12.5) | 1,520 (4.0) |
| Tumor that may influence HCG | 1,769 (0.7) | 951 (0.6) | 44 (0.4) | 50 (0.2) | 98 (0.3) | 5 (0.2) | 304 (9.7) | 20 (1.0) | 19 (0.5) | 278 (0.7) |

[a] Within 270 days after the cohort entry date
[b] ICD9 codes 630, 631 (i.e. Hydatidiform mole or other abnormal products of conception)
[c] Quantitative lab considered positive if >25 mIU/mL

The subset of women with only a negative HCG test result (8.0% of the cohort; n = 21,368) by definition entered the cohort as a result of either a prenatal diagnosis or procedure code (n = 10,858) or a coded pregnancy outcome (n = 10,510). Among the 10,858 women with a prenatal diagnosis or procedure code in the database, the most common initial codes were normal pregnancy, ultrasound of pregnant uterus, and hemorrhage in early pregnancy. These women often had a negative HCG result value with a prenatal diagnosis coded up to a few weeks later (and no additional HCG results). In a few situations, the coded prenatal diagnosis or procedure was followed by a negative qualitative or low (< 25 mIU/mL) quantitative HCG result value. Also, a higher proportion of women with extra-uterine pregnancies only had negative HCG result values (27.2% of the extra-uterine pregnancies versus 8.0% of the cohort overall). The proportion of this subset of women with a pregnancy that ended in a full term or preterm single live birth was similar to that of the cohort overall (61.3% in the subset; 62.4% in the full cohort; not shown in table). A greater proportion of this 8.0% subset of women had no end of pregnancy indicator in the database (21.9% versus 14.3% in the full cohort).

The subset of women with a pregnancy identified only by a positive HCG result value (5.1%; n = 13,665) and no prenatal diagnosis or procedure code (74.7% had a qualitative HCG result; 33.7% had a quantitative HCG result) had end of pregnancy indicators that differed dramatically from those of the full cohort (Table 22). Only 1.0% (n = 135) had a pregnancy that ended in a full term or preterm single live birth. 50.9% ended in abortion, 22.1% ended in miscarriage, and 22.5% had no end of pregnancy indicator found. As shown in Table 22, these women's end of pregnancy indicators varied noticeably by site (e.g., abortions: 54.7% at site 2, 0.7% at site 3; no end of pregnancy indicator found: 20.4% at site 2, 52.0% at site 3). The demographics and other characteristics of this subset of women are also in Table 22.

**Table 22. Characteristics of the 5.1% Subset of the Pregnancy Cohort Identified by Positive Lab Result Only, Overall and by Site**

| Characteristic | Total, N (%) | Data Partner Site, N (%) | | |
|---|---|---|---|---|
| | | Site 1 | Site 2 | Site 3 |
| Number of women | 13,665 | 679 | 12,582 | 404 |
| End of pregnancy indictor [b] | | | | |
| Single live birth: Term | 28 (0.2) | 6 (0.9) | 18 (0.1) | 4 (1.0) |
| Single live birth: Preterm | 107 (0.8) | 7 (1.0) | 97 (0.8) | 3 (0.7) |
| Abortion | 6,953 (50.9) | 73 (10.8) | 6,877 (54.7) | 3 (0.7) |
| Miscarriage | 3,014 (22.1) | 226 (33.3) | 2,649 (21.1) | 139 (34.4) |
| Fetal death | 16 (0.1) | 0 (0.0) | 16 (0.1) | 0 (0.0) |
| Extra-uterine | 274 (2.0) | 27 (4.0) | 215 (1.7) | 32 (7.9) |
| Abnormal conception product [c] | 192 (1.4) | 32 (4.7) | 147 (1.2) | 13 (3.2) |
| Multiple births | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| No end of pregnancy indicator found | 3,081 (22.5) | 308 (45.4) | 2,563 (20.4) | 210 (52.0) |
| Reason end of pregnancy indicator not found | | | | |
| Death | 2 (0.0) | 0 (0.0) | 2 (0.0) | 0 (0.0) |
| End of study (estimated due date beyond 12/31/2013) | 613 (4.5) | 65 (9.6) | 511 (4.1) | 37 (9.2) |
| Unknown | 2,466 (18.0) | 243 (35.8) | 2,050 (16.3) | 173 (42.8) |
| Age in years, mean (SD) | 29.0 (7.8) | 29.0 (7.5) | 28.9 (7.8) | 31.8 (7.2) |
| Race | | | | |
| White | 5,311 (38.9) | 408 (60.1) | 4,882 (38.8) | 21 (5.2) |
| Black | 2,288 (16.7) | 57 (8.4) | 2,225 (17.7) | 6 (1.5) |
| Other | 3,012 (22.0) | 36 (5.3) | 2,975 (23.6) | 1 (0.2) |
| Unknown | 3,054 (22.3) | 178 (26.2) | 2,500 (19.9) | 376 (93.1) |
| Hispanic ethnicity (Y vs N/unknown) | 2,948 (21.6) | 134 (19.7) | 2,812 (22.3) | 2 (0.5) |
| Qualitative or Quantitative HCG lab result value [d] | | | | |
| No HCG lab result | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| Only negative HCG lab result | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| 1+ Positive HCG lab result; women with lab result and no pregnancy diagnosis | 13,665 (100.0) | 679 (100.0) | 12,582 (100.0) | 404 (100.0) |
| 1+ Positive HCG lab result; women with diagnosis and lab result | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| Qualitative HCG lab | | | | |
| No Qualitative HCG lab result | 2,986 (21.9) | 180 (26.5) | 2,514 (20.0) | 292 (72.3) |
| Only negative Qualitative HCG lab result | 466 (3.4) | 79 (11.6) | 376 (3.0) | 11 (2.7) |
| 1+ Positive Qualitative HCG lab; women with lab result and no pregnancy diagnosis | 10,213 (74.7) | 420 (61.9) | 9,692 (77.0) | 101 (25.0) |
| 1+ Positive Qualitative HCG lab; women with pregnancy diagnosis and lab result | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| Quantitative HCG lab [e] | | | | |
| No Quantitative HCG lab result | 8,635 (63.2) | 293 (43.2) | 8,271 (65.7) | 71 (17.6) |
| Only negative Quantitative HCG lab result | 426 (3.1) | 25 (3.7) | 390 (3.1) | 11 (2.7) |
| 1+ Positive Quantitative HCG lab; women with lab results and no pregnancy diagnosis | 4,604 (33.7) | 361 (53.2) | 3,921 (31.2) | 322 (79.7) |
| 1+ Positive Quantitative HCG lab result; women with pregnancy diagnosis and lab result | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| Medical Utilization [d] | | | | |
| No encounters prior to first pregnancy indicator | 11 (0.1) | 1 (0.1) | 4 (0.0) | 6 (1.5) |
| Ambulatory visits, mean (SD) | 6.1 (7.5) | 8.3 (8.7) | 5.9 (7.4) | 7.3 (7.4) |
| Emergency department visits, % yes | 2,023 (14.8) | 91 (13.4) | 1,882 (15.0) | 50 (12.4) |

| Characteristic | Total, N (%) | Data Partner Site, N (%) | | |
|---|---|---|---|---|
| | | Site 1 | Site 2 | Site 3 |
| Hospitalization, % yes | 217 (1.6) | 9 (1.3) | 196 (1.6) | 12 (3.0) |
| Institutional stay, % yes | 19 (0.1) | 0 (0.0) | 5 (0.0) | 14 (3.5) |
| HCG injections | 319 (2.3) | 40 (5.9) | 277 (2.2) | 2 (0.5) |
| Tumor that may influence HCG | 28 (0.2) | 2 (0.3) | 23 (0.2) | 3 (0.7) |

[a] Women included in the cohort have enrollment in the health plan for at least 183 days prior to the cohort entry date and 270 days after the cohort entry date. Cohort entry date is the date of first indicator of pregnancy (i.e., diagnosis code, procedure code, or laboratory test result)
[b] Within 270 days after the cohort entry date
[c] ICD9 codes 630, 631 (i.e. Hydatidiform mole or other abnormal products of conception)
[d] Quantitative lab considered positive if >25 mIU/mL
[e] During 183 days prior to cohort entry (earliest pregnancy indicator)

Included among the 47.6% of the cohort with no HCG laboratory result (N = 127,613, Tables 20 and 21) were a subset of women whose pregnancy was not identified until there was a coded end of pregnancy indictor in the database (Table 23; 8.9% of the cohort, N = 23,741). These women had term or preterm single live births at a proportion similar to the full cohort (59.7% in this subset versus 62.4% in the full cohort). However, 31.5% of the pregnancies in this subset of women ended in abortion or miscarriage (19.1% in the full cohort) and 6.9% of these pregnancies were extra-uterine (1.2% in the full cohort).

**Table 23. End of Pregnancy Indicator for the 8.9% Subset of the Pregnancy Cohort with No Prior Pregnancy Indicator, Overall and by Site**

| End of Pregnancy Indicator | Total, N (%) | Data Partner Site, N (%) | | |
|---|---|---|---|---|
| | | Site 1 | Site 2 | Site 3 |
| Number of women | 23,741 | 1,922 | 18,092 | 3,727 |
| Single live birth: Term | 13,561 (57.1) | 931 (48.4) | 11,543 (63.8) | 1,087 (29.2) |
| Single live birth: Preterm | 616 (2.6) | 119 (6.2) | 357 (2.0) | 140 (3.8) |
| Abortion | 2,467 (10.4) | 216 (11.2) | 2,190 (12.1) | 61 (1.6) |
| Miscarriage | 4,998 (21.1) | 528 (27.5) | 2,591 (14.3) | 1,879 (50.4) |
| Fetal death | 65 (0.3) | 8 (0.4) | 39 (0.2) | 18 (0.5) |
| Extra-uterine | 1,636 (6.9) | 96 (5.0) | 1,218 (6.7) | 322 (8.6) |
| Abnormal conception product | 308 (1.3) | 12 (0.6) | 111 (0.6) | 185 (5.0) |
| Multiple births | 90 (0.4) | 12 (0.6) | 43 (0.2) | 35 (0.9) |
| No end of pregnancy indicator found | 0 | 0 | 0 | 0 |

### c. Cohort Identification Test Case 1 results and discussion

***Pregnancies gained by including those identified using laboratory results only (research question 1)***

For 5.1% of the cohort (N = 13,665), the only indicator of the pregnancy in the database was a positive qualitative or quantitative HCG result (i.e., no pregnancy-related diagnosis or procedure code). This ranged from 0.7% at site 3 to 6.7% at site 2 (Tables 20 and 24). An additional 39.4% (N = 105,573) of women had both a positive HCG result and a prenatal diagnosis or procedure code; 46.7% (N = 125,240) of women had only a prenatal diagnosis or procedure code.

For 8.9% (N = 23,741) of the cohort, the pregnancy was not documented until the pregnancy outcome was coded (Table 24). This subset includes many women without prenatal care. It also includes women for whom data from the entire pregnancy was not available (N = 2012, 8.5% [of 23,741]; site variability within this subset: 7.8% at site 1, 4.4% at site 2, 28.6% at site 3). That is, women were required to have a minimum of 183 days of enrollment prior to cohort entry, but for 2012 women, their enrollment length was between 183 and 270 days prior to pregnancy outcome, so their enrollment likely did not extend

back in time far enough to include early codes or HCG results that might have been available, particularly for term pregnancies. However, this does not mitigate the fact that these women did not have any other evidence of their pregnancies during their enrollment within the 183 days prior to their pregnancy outcomes.

**Table 24. Pregnancy Identification by Pregnancy Indicator, Overall and by Site**

| First Pregnancy Indicator | Total N =268,219 (%) | Data Partner Site | | |
|---|---|---|---|---|
| | | Site 1 N = 24,649 (%) | Site 2 N = 187,471 (%) | Site 3 N = 56,099 (%) |
| Prenatal diagnosis or procedure only | 125,240 (46.7) | 7,515 (30.5) | 76,554 (40.8) | 41,171 (73.4) |
| Prenatal positive lab result only | 13,665 (5.1) | 679 (2.8) | 12,582 (6.7) | 404 (0.7) |
| Prenatal diagnosis or procedure and positive lab result | 105,573 (39.4) | 14,533 (59.0) | 80,243 (42.8) | 10,797 (19.2) |
| End of pregnancy indicator only | 23,741 (8.9) | 1,922 (7.8) | 18,092 (9.7) | 3,727 (6.6) |

Table 25 presents the number and percent of each end of pregnancy indicator type that were identified by HCG results only. The percent of pregnancies indicated by only an HCG result varied dramatically across end of pregnancy indicator types, ranging from 0% among single live term and multiple births, to 32.6% among pregnancies that ended with abortion.

**Table 25. Number and Percent of Pregnancies Identified by HCG Laboratory Test Results Only, Stratified by End of Pregnancy Indicator, Overall and by Site**

| End of Pregnancy Indicator | Number with Prenatal Positive Lab Result Only divided by Total Number with that End of Pregnancy Indicator (%) | | | |
|---|---|---|---|---|
| | Total (%) | Data Partner Site, N (%) | | |
| | | Site 1 | Site 2 | Site 3 |
| Term single live births | 28/157,076 (0.0) | 6/15,060 (0.0) | 18/107,675 (0.0) | 4/34,341 (0.0) |
| Preterm single live births | 107/10,261 (1.0) | 7/974 (0.7) | 97/6,796 (1.4) | 3/2,491 (0.1) |
| Abortions | 6,953/21,357 (32.6) | 73/1,127 (6.5) | 6,877/19,991 (34.4) | 3/239 (1.3) |
| Miscarriages | 3,014/29,899 (10.1) | 226/2,637 (8.6) | 2,649/20,696 (12.8) | 139/6,566 (2.1) |
| Fetal deaths | 16/2,006 (0.8) | 0/214 (0.0) | 16/1,440 (1.1) | 0/352 (0.0) |
| Extra-uterine pregnancies | 274/3,146 (8.7) | 27/240 (11.3) | 215/2,127 (10.1) | 32/779 (4.1) |
| Abnormal product of conception | 192/2,083 (9.2) | 32/195 (16.4) | 147/1,252 (11.7) | 13/636 (2.0) |
| Multiple births | 0/4,027 (0.0) | 0/386 (0.0) | 0/2,614 (0.0) | 0/1,027 (0.0) |
| End of pregnancy indicator not found | 3,081/38,364 (8.0) | 308/3,816 (8.1) | 2.563/24,880 (10.3) | 210/9,668 (2.2) |
| Total | 13,665 | 679 | 12,582 | 404 |

***How the first positive laboratory result changes the timing of when the pregnancy was first identified in the electronic data***

Table 26 indicates that, in the electronic database, among women with a positive HCG result, 60.7% (n = 39,559) had their pregnancy identified earlier using the HCG result than using the diagnosis or procedure code (range across sites: 19.3% at site 3 to 73.9% at site 2). For 19.6% (n = 12,742) of women, the first positive HCG result and the prenatal diagnosis or procedure code occurred on the same date (range across sites: 14.9% at site 1 to 22.5% at site 3). For 19.7% (n = 12,830), the positive HCG result date was later than the date of the coded prenatal diagnosis or procedure (range across sites: 6.0% at site 2 to 58.2% at site 3).

**Table 26. First Pregnancy Indicator among Pregnancies with Prenatal Positive Lab Result and Prenatal Diagnosis or Procedure Code and Single Live Birth Diagnosis or Procedure Code, Overall and by Site**

| Pregnancy Indicator | Total, N = 65,131 (%) | Data Partner Site | | |
|---|---|---|---|---|
| | | Site 1 N =10,395 (%) | Site 2 N = 46,824 (%) | Site 3 N = 7,912 (%) |
| Positive lab result occurs before diagnosis/procedure code | 39,559 (60.7) | 3,433 (33.0) | 34,596 (73.9) | 1,530 (19.3) |
| Positive lab result occurs on the same day as the diagnosis/procedure code | 12,742 (19.6) | 1,548 (14.9) | 9,413 (20.1) | 1,781 (22.5) |
| Positive lab result occurs after diagnosis/procedure code | 12,830 (19.7) | 5,414 (52.1) | 2,815 (6.0) | 4,601 (58.2) |

For the subset of women with live births whose first pregnancy indicators were both a prenatal coded diagnosis or procedure and HCG results, the first positive HCG result was documented in the electronic database a mean (SD) of 15.7 (12.2) days before the prenatal diagnosis or procedure code (Table 27). For women with abortion or miscarriage, the mean time the positive HCG result was recorded in the electronic database was 11.6 (15.7) days prior to the prenatal diagnosis or procedure code (Table 27).

**Table 27. Lead Time of Positive HCG Laboratory Test Result Prior to Prenatal Diagnosis or Procedure Codes among Pregnancies with a Live Birth, Abortion, or Miscarriage that had Both Prenatal Diagnosis or Procedure Codes and Laboratory Test Results, Overall and by Site**

| Characteristic | Total, N (%) | Data Partner Site, N (%) | | |
|---|---|---|---|---|
| | | Site 1 | Site 2 | Site 3 |
| Single live birth: Term or Preterm, N | 39,559 | 3,433 | 34,596 | 1,530 |
| Days lead time, mean (SD) | 15.7 (12.2) | 16.9 (11.8) | 15.6 (12.0) | 17.4 (17.2) |
| Pregnancy loss: Abortion or Miscarriage, N | 13,655 | 848 | 12,490 | 317 |
| Days lead time, mean (SD) | 11.6 (15.7) | 13.6 (20.2) | 11.6 (15.5) | 9.6 (9.0) |

For women where the first indicator of pregnancy was the HCG result in the distributed database, the vast majority (n = 51,440, 98.1%) of estimated gestational periods fell within the 270 day gestational age metric applied in previous observation studies that did not consider laboratory indicators of pregnancy (metric: delivery code date minus 270 days equals estimated length of gestation [93-95] )(Table 28). The small proportion of pregnancies with a positive HCG result outside the estimated gestational period were most often associated with a conflicting coded diagnoses (e.g., a coded live birth and a coded miscarriage that appeared to apply to the same pregnancy episode; Table 28).

**Table 28. Positive HCG Laboratory Test Result within Estimated Gestation Period a among Pregnancies with Prenatal Positive HCG Laboratory Test Result Only or with both a Positive HCG Laboratory Test Result Prior to a Prenatal Diagnosis or Procedure Code and a Single Live Birth Diagnosis or Procedure Code, Overall and by Site**

| Characteristic | Total N = 52,436 (%) | Data Partner Site | | |
|---|---|---|---|---|
| | | Site 1 N = 4,994 (%) | Site 2 N = 44,124 (%) | Site 3 N = 3,318 (%) |
| Positive lab within estimated gestation period | 51,440 (98.1) | 4,915 (98.4) | 43,287 (98.1) | 3,238 (97.6) |
| Positive lab outside estimated gestation period | 996 (1.9) | 79 (1.6) | 837 (1.9) | 80 (2.4) |
| Conflicting end of pregnancy diagnosis or procedure code in addition to live birth code [b] | 871 (87.4) | 67 (84.8) | 730 (87.2) | 74 (92.5) |
| Abortion | 270 (31.0) | 5 (7.5) | 264 (36.2) | 1 (1.4) |
| Miscarriage | 412 (47.3) | 49 (73.1) | 331 (45.3) | 32 (43.2) |
| Fetal death | 46 (5.3) | 7 (10.4) | 34 (4.7) | 5 (6.8) |
| Extra-uterine | 88 (10.1) | 1 (1.5) | 55 (7.5) | 32 (43.2) |
| Abnormal conception product | 55 (6.3) | 5 (7.5) | 46 (6.3) | 4 (5.4) |
| Indication of preterm delivery conflicts with observed length of pregnancy | 125 (12.6) | 12 (15.2) | 107 (12.8) | 6 (7.5) |

[a] Estimated gestational age determined according to preterm diagnosis codes (if given) or set to 270 days for full term pregnancies

[b] N = 53 conflicting end of pregnancy codes (in addition to the live birth code) were coded as both miscarriage and abortion. These are included in the abortion row. Other conflicting end of pregnancy codes that were coded at both fetal death and abortion or miscarriage (N = 7), abnormal conception product and abortion or miscarriage (N = 2), or extra-uterine and miscarriage or abortion (N = 2) are included in the fetal death, abnormal conception product, or extra-uterine rows. All other conflicting end of pregnancy codes only had one conflicting code listed.

Considering the initial 90 days of the 270 day metric as the first trimester, among women with single live births who had the pregnancy first identified using a HCG result (with or without a diagnosis or procedure code), 79.8% (n = 52,103) had both the first laboratory result and the first diagnosis or procedure code recorded during the estimated first trimester (Table 29). Only 5.5% (n = 3583) had neither a diagnosis or procedure code nor an HCG result first recorded in the estimated first trimester.

**Table 29. Proportion of Single Live Births with Pregnancy Indicator Recorded within First Trimester Based on Estimated Gestational Period a among Pregnancies with Prenatal Positive HCG Laboratory Test, Overall and by Site**

| Characteristic | Total N = 65,266 (%) | Data Partner Site | | |
|---|---|---|---|---|
| | | Site 1 N = 10,408 (%) | Site 2 N = 46,939 (%) | Site 3 N = 7,919 (%) |
| Neither diagnosis/procedure nor lab result in 1st trimester | 3,583 (5.5) | 770 (7.4) | 2,095 (4.5) | 718 (9.1) |
| Lab only in 1st trimester | 854 (1.3) | 77 (0.7) | 694 (1.5) | 83 (1.0) |
| Diagnosis/procedure only in 1st trimester | 8,726 (13.4) | 4,835 (46.5) | 569 (1.2) | 3,322 (41.9) |
| Diagnosis/procedure and lab result in 1st trimester | 52,103 (79.8) | 4,726 (45.4) | 43,581 (92.8) | 3,796 (47.9) |

[a] Estimated gestational age determined according to preterm diagnosis codes (if given) or set to 270 days for full term pregnancies

### d. Summary of Cohort Identification Test Case 1

In summary, this analysis estimated that using positive HCG laboratory results increased the number of women identified as pregnant (during the prenatal period) by 5.1% (n = 13,665) compared to using diagnosis and procedure codes alone; this varied across sites even within similarly-integrated healthcare delivery system sites (site 1 = 2.8%, site 2 = 6.7%). Among 268,219 women in the MSDD from these three sites who were pregnant between 2008 and 2013, 52.4% had at least one positive HCG result recorded, with availability of HCG results varying by site. For the 5.1% of the cohort whose only indicator of pregnancy in the dataset was the positive HCG result, 8.0% (N = 3081) had no end of pregnancy indicator. These women could have obtained abortions outside of their healthcare insurer, miscarried without seeking medical care, or delivered a live infant at home. Some women without end of pregnancy indicators could be women who sought abortions because, for 32.6% of the women whose pregnancies were known to end in abortion, the only indicator of pregnancy was a positive HCG laboratory test result. For this 5.1%, the incidence of miscarriage was also much higher than for the full cohort. For studying drug safety in pregnancy, identifying pregnancies based on HCG results might be particularly important if use of the drug of interest could be associated with miscarriage. In future work it will be important to study these subgroups of women better, in particular within the context of risks of drug exposures during pregnancy.

For women with live births, this analysis suggests using positive HCG results would enable cohort identification more than two weeks (mean 15.7 days) earlier than using prenatal diagnosis or procedure codes alone. Women with abortion or miscarriage also had the positive HCG result recorded an average 11.6 days before the prenatal diagnosis or procedure code.

Nearly all (98.1%) positive HCG results fell within the 270 day gestational age metric, suggesting that continued use of the metric in observational database studies in determining gestational age from live births is appropriate.

## 2. Cohort Identification Test Case 2: enhanced identification of a cohort of adults with chronic kidney disease (CKD) through use of serum creatinine laboratory results data to estimate patients' glomerular filtration rate

### a. Cohort Identification Test Case 2 cohort development

This test case was focused on a cohort of individuals with one or more indicators of chronic kidney disease (CKD). Primary analyses were completed using a 2012 cohort; selected analyses were completed for 2010 and 2008 cohorts. In this test case, we examined whether estimating the glomerular filtration rate (eGFR)[2] using outpatient serum creatinine results values applied to the CKD Epidemiology Collaboration (CKD-EPI) equation [96] (Figure 5) augmented identifying a cohort of adults with CKD aged ≥ 21 through ≤ 89 years. The Workgroup also assessed the percent agreement between using serum creatinine laboratory test result values to estimate eGFR and coded diagnosis in this patient cohort.

**Figure 5. The 2009 CKD-EPI Creatinine Equation [96]**

$$GFR = 141 \times \min(Scr/\kappa, 1)^{\alpha} \times \max(Scr/\kappa, 1)^{-1.209} \times 0.993^{Age} \times 1.018[\text{if female}] \times 1.159[\text{if black}]$$

$\kappa = 0.7$ if female
$\kappa = 0.9$ if male

$\alpha = -0.329$ if female
$\alpha = -0.411$ if male

min = The minimum of $Scr/\kappa$ or 1
max = The maximum of $Scr/\kappa$ or 1

Scr = serum creatinine (mg/dL)

Similar to the previous cohort identification test case, because the primary purpose was to determine whether considering laboratory results enhanced cohort identification, no medical product exposure was required. Rather, the "exposure" was the availability of the creatinine result value (used to estimate GFR). Lists of codes used to identify patients with CKD and of codes used to exclude patients with kidney transplant or dialysis at baseline are provided in Appendix D. eGFR was defined as compatible with CKD Stage 3 or higher when the eGFR value was < 60 ml/min/1.73m$^2$.

This test case was expected to identify a mixed cohort of patients with prevalent and incident CKD because we only required 183 days minimum enrollment prior to the first indicator of CKD in 2012. This design was also expected to include a higher proportion of patients with Stages 4 and 5 CKD. Identifying and including patients with Stages 4 and 5 CKD is important for drug safety research because later stage CKD patients are more often candidates for medication dosage or frequency of medication dosing adjustment and are at higher risk of adverse outcomes if medications are not appropriately adjusted for level of renal dysfunction.

[2] We quantified the n and % of patients with and without race data in this cohort. For patients without race data, we assumed non-African American when employing the CKD-EPI equation to estimate GFR.

The following research questions were addressed using the CKD cohort identified in this test case:

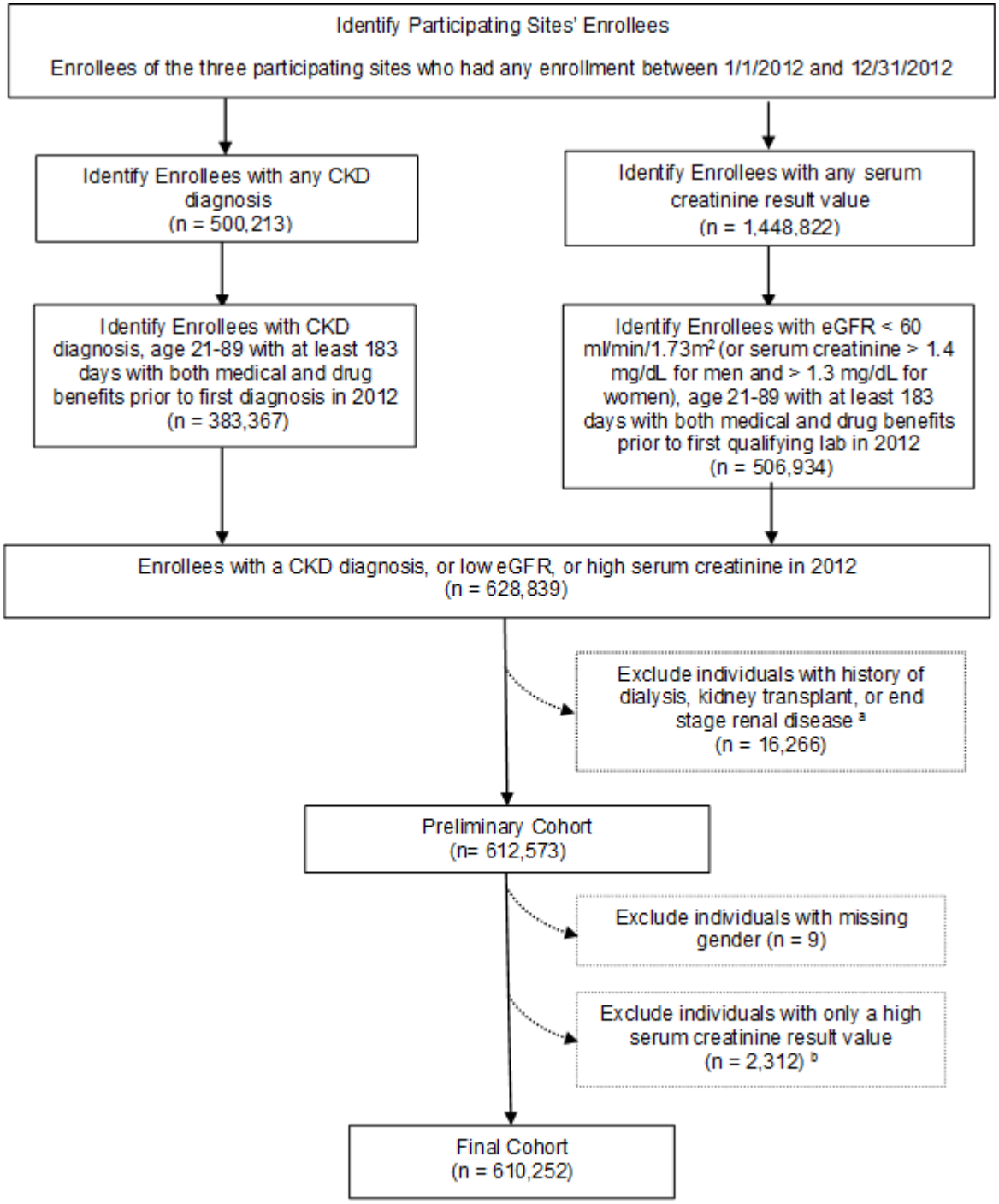
1. What is the percent agreement between CKD identified using laboratory result values (at least 2 eGFR values <60ml/min/1.73m$^2$ with no intervening values $\geq$ 60 measured at least 90 days apart) compared to at least 1 coded CKD diagnosis?
2. What is the percent agreement between CKD identified using at least 2 eGFR values <60ml/min/1.73m$^2$ (with no intervening values $\geq$ 60) measured at least 90 days apart compared to identifying CKD using at least 2 coded diagnoses of CKD?
3. What is the percent agreement between CKD identified using at least 1 eGFR value <60ml/min/1.73m$^2$ compared to identifying CKD using at least 1 coded diagnosis of CKD.
4. Describe the CKD cohort by age category < 65, $\geq$ 65 - < 75, and $\geq$ 75 – 89 years.
5. Describe the CKD cohort by CKD Stage 3, Stage 4, Stage 5, and not staged.
6. Describe the numbers and magnitude of increase in cohort size by calendar year. The primary cohort for questions 1 – 5 is from 2012. Cohorts from 2010 and 2008 were also identified and employed to answer this question.
7. Each analysis was repeated stratified by Data Partner.

These analyses answer questions about how many additional patients are gained in a CKD cohort by including laboratory results and assist in understanding whether differences in availability of laboratory result values results in variation in cohort characteristics.

The *diagnosis* index date ($T_0$) was defined as the first CKD diagnosis code in 2012. The *laboratory result* index date ($T_0L$) was defined as the first eGFR (calculated from serum creatinine result value) < 60 ml/min/1.73m$^2$ when that serum creatinine result date preceded the date of the first CKD diagnosis or when there was no CKD diagnosis for a patient. The earliest of $T_0$ or $T_0L$ was used for baseline medical and drug coverage determination. Medical and drug coverage was required for $\geq$ 183 days prior to $T_0$ or $T_0L$ in 2012. Similarly, for estimating whether cohort augmentation changed over time, medical and drug coverage was required for $\geq$ 183 days prior to index date for 2008 and 2010. For the 2012 cohort, the first diagnosis or creatinine result value must have been in 2012 and the second diagnosis or creatinine result value was identified within the 365 days following $T_0$ or $T_0L$. A similar approach was used for the 2010 and 2008 cohorts. Thus, a 365-day timeframe was used to identify all diagnosis or laboratory-based indicators of CKD for each patient. Figure 6 shows the preliminary cohort (N = 612,573). After removing individuals who had an elevated creatinine but whose eGFR was > 60 ml/min/1.73m$^2$ and those for whom no gender was available, the final CKD cohort included 610,252 individuals (Figure 6). Further details on the specifications of the test case, such as covariate data collected, are in Appendix D.

No outcome was required for this test case. We applied these censoring rules: disenrollment from the health plan, dialysis, death, day 365 after cohort entry, and end of study timeframe.

**Figure 6. Test Case for Cohort Identification Test Case 2, Enhanced Identification of a Cohort of Adults with Chronic Kidney Disease (CKD) through Use of Serum Creatinine Laboratory Results Values to Estimate Glomerular Filtration Rate**




[a] Codes used to identify individuals to exclude are in Appendix D

[b] Although these men had serum creatinine > 1.4 mg/dL and these women had serum creatinine > 1.3 mg/dL, their calculated eGFR was > 60 ml/min/1.73m²

### b. Cohort Identification Test Case 2 descriptive analysis (research questions 4 and 5)

A detailed description of the 2012 CKD cohort is in Table 30. To describe the CKD cohort and align with the primary purpose of determining whether laboratory results enhanced cohort identification, if individuals had at least two coded CKD diagnoses within 365 days, they were preferentially assigned to that subgroup whether or not they also had eGFR estimate(s) available. Overall, 299,751 of the 610,252 (49.1%) individuals in the cohort had at least two CKD diagnoses within 365 days. The proportion of the cohort with at least two CKD diagnoses at each site did not vary dramatically from the proportion of the cohort contributed by that site (Table 30). However, estimates within site varied from 37.5% of the cohort from site 1 (i.e., 17,593 of 46,886; site details in Appendix H) to 50.0% and 50.3% of the cohort from sites 3 and 2, respectively.

If the CKD definition was relaxed to include patients with only one coded CKD diagnosis and one or more eGFR < 60 ml/min/1.73m$^2$ within 365 days, an additional 50,540 (8.3%) individuals were identified. If the definition was expanded to include patients with at least two eGFR within 365 days of each other with no coded diagnosis, an additional 77,855 (12.8%) individuals were included. If broader definitions were applied such as requiring only one diagnosis or only one low eGFR, an additional 31,948 (5.2%) and 150,158 (24.6%) of individuals were included, respectively. Requiring only one low eGFR within 365 days had the greatest variability in cohort inclusion across sites ranging from 21.8% at site 3 to 38.9% at site 1. The proportions with at least one eGFR plus one diagnosis code (range: 7.6% to 9.7%), and with two or more low eGFR (range: 10.6% to 13.8%) did not vary dramatically by site. This lesser variability in site-specific cohort proportions with available laboratory results differs from the site variations found in other test cases in this project. Additional information about the characteristics of the site-specific cohorts is in Appendix H.

**Table 30. Characteristics of Individuals in the 2012 Chronic Kidney Disease Overall Cohort Identification Test Case 2 Population**

| Characteristics [a] | CKD Identified by ≥ 1 Diagnosis Code, by ≥ 1 eGFR values <60 ml/min/1.73m$^2$ (Calculated from Serum Creatinine Result Values), or by Both Methods | | | | | |
|---|---|---|---|---|---|---|
| | ≥ 2 CKD Diagnosis Codes [b] N = 299,751 (49.1%) | 1 CKD Diagnosis Code and ≥ 1 low eGFR N = 50,540 (8.3%) | ≥ 2 Low eGFRs (no diagnosis) N = 77,855 (12.8%) | 1 CKD Diagnosis only N = 31,948 (5.2%) | 1 low eGFR only N = 150,158 (24.6 %) | Total N = 610,252 |
| CKD stage from diagnosis code, N [c] | 299,751 | 50,540 | N/A | 31,948 | N/A | 382,239 |
| Stage 3 | 198,545 (66.2) | 21,274 (42.1) | | 12,532 (39.2) | | 232,351 (60.8) |
| Stage 4 | 16,490 (5.5) | 599 (1.2) | | 431 (1.3) | | 17,520 (4.6) |
| Stage 5 | 1,272 (0.4) | 49 (0.1) | | 125 (0.4) | | 1,446 (0.4) |
| Stage unspecified/other | 83,444 (27.8) | 28,618 (56.6) | | 18,860 (59.0) | | 130,922 (34.3) |
| Estimated CKD stage from eGFR, N | 220,113 | 50,540 | 77,855 | N/A | 150,158 | 498,666 |
| Stage 3 (30-59 ml/min/1.73m$^2$) | 189,776 (86.2) | 47,434 (93.9) | 76,546 (98.3) | | 148,218 (98.7) | 461,974 (92.6) |
| Stage 4 (15-29 ml/min/1.73m$^2$) | 26,375 (12.0) | 2,636 (5.2) | 1,274 (1.6) | | 1,622 (1.1) | 31,907 (6.4) |
| Stage 5 (< 15 ml/min/1.73m$^2$) | 3,962 (1.8) | 470 (0.9) | 35 (0.0) | | 318 (0.2) | 4,785 (1.0) |

| Characteristics [a] | CKD Identified by ≥ 1 Diagnosis Code, by ≥ 1 eGFR values <60 ml/min/1.73m² (Calculated from Serum Creatinine Result Values), or by Both Methods | | | | | |
|---|---|---|---|---|---|---|
| | ≥ 2 CKD Diagnosis Codes [b] N = 299,751 (49.1%) | 1 CKD Diagnosis Code and ≥ 1 low eGFR N = 50,540 (8.3%) | ≥ 2 Low eGFRs (no diagnosis) N = 77,855 (12.8%) | 1 CKD Diagnosis only N = 31,948 (5.2%) | 1 low eGFR only N = 150,158 (24.6 %) | Total N = 610,252 |
| Estimated CKD stage from diagnosis and eGFR [d] | 299,751 | 54,540 | 77,855 | 31,948 | 150,158 | 610,252 |
| Stage 3 (30-59 ml/min/1.73m²) | 228,144 (76.1) | 47,048 (93.1) | 76,546 (98.3) | 12,532 (39.2) | 148,218 (98.7) | 512,488 (84.0) |
| Stage 4 (15-29 ml/min/1.73m²) | 33,737 (11.3) | 2,990 (5.9) | 1,274 (1.6) | 431 (1.3) | 1,622 (1.1) | 40,054 (6.6) |
| Stage 5 (< 15 ml/min/1.73m²) | 4,692 (1.6) | 502 (1.0) | 35 (0.0) | 125 (0.4) | 318 (0.2) | 5,672 (0.9) |
| Stage unspecified/other | 33,178 (11.1) | 0 (0.0) | 0 (0.0) | 18,860 (59.0) | 0 (0.0) | 52,038 (8.5) |
| Age in years, mean (SD) | 73.9 (9.6) | 73.2 (10.2) | 74.4 (8.6) | 70.4 (11.4) | 71.1 (10.4) | 73.1 (9.9) |
| Age categories, years<br>< 65<br>65-74<br>75-89 | <br>40,309 (13.4)<br>104,372 (34.8)<br>155,070 (51.7) | <br>8,383 (16.6)<br>17,341 (34.2)<br>24,816 (49.1) | <br>9,085 (11.7)<br>28,766 (36.9)<br>40,004 (49.1) | <br>7,332 (22.9)<br>12,354 (38.7)<br>12,262 (38.4) | <br>33,690 (22.4)<br>56,795 (37.8)<br>59,673 (39.7) | <br>98,799 (16.2)<br>219,628 (36.0)<br>291,825 (47.8) |
| Female sex | 157,478 (52.5) | 27,964 (55.3) | 49,669 (63.8) | 15,701 (49.1) | 90,457 (60.2) | 341,269 (55.9) |
| Site | | | | | | |
| 1 | 17,593 (5.9) | 3,982 (7.9) | 5,985 (7.7) | 1,105 (3.5) | 18,221 (12.1) | 46,886 (7.7) |
| 2 | 93,350 (31.1) | 17,976 (35.6) | 19,574 (25.1) | 5,180 (16.2) | 49,484 (33.0) | 185,564 (30.4) |
| 3 | 188,808 (63.0) | 28,582 (56.6) | 52,296 (67.2) | 25,663 (80.3) | 82,453 (54.9) | 377,802 (61.9) |
| Any serum creatinine value available in 2012 | 250,927 (83.7) | 49,896 (98.7)[d] | 77,855 (100.0) | 16,234 (50.8) | 150,158 (100) | 545,069 (89.3) |
| Serum creatinine procedure code in 2012 | 291,357 (97.4) | 49,900 (98.8) | 77,240 (99.3) | 28,447 (90.3) | 148,693 (99.2) | 595,637 (97.9) |
| Race | | | | | | |
| White | 222,415 (74.2) | 39,291 (77.7) | 64,334 (82.6) | 23,122 (72.4) | 118,837 (79.1) | 467,999 (76.7) |
| Black | 40,875 (13.6) | 5,130 (10.2) | 5,258 (6.8) | 4,514 (14.1) | 10,376 (6.9) | 66,153 (10.8) |
| Other | 19,330 (6.4) | 2,986 (5.9) | 3,773 (4.8) | 1,446 (4.5) | 8,270 (5.5) | 35,805 (5.9) |
| Unknown | 17,131 (5.7) | 3,133 (6.2) | 4,490 (5.8) | 2,866 (9.0) | 12,675 (8.4) | 40,295 (6.6) |
| No encounters in prior 183 days | 14,049 (4.7) | 2,983 (5.9) | 3,790 (4.9) | 1,458 (4.6) | 9,209 (6.1) | 31,489 (5.2) |
| Number of ambulatory medical visits during baseline, mean (SD) | 7.9 (8.5) | 7.1 (7.9) | 6.2 (6.4) | 8.1 (8.7) | 6.2 (7.1) | 7.2 (7.9) |
| Emergency department visit during baseline, N (%) yes | 34,224 (11.4) | 6,087 (12.0) | 7,063 (9.1) | 3,945 (12.3) | 16,518 (11.0) | 67,837 (11.1) |
| Hospitalization during baseline, N (%) yes | 25,839 (8.6) | 4,411 (8.7) | 4,110 (5.3) | 3,118 (9.8) | 11,137 (7.4) | 48,615 (8.0) |

| Characteristics [a] | CKD Identified by ≥ 1 Diagnosis Code, by ≥ 1 eGFR values <60 ml/min/1.73m² (Calculated from Serum Creatinine Result Values), or by Both Methods | | | | | |
|---|---|---|---|---|---|---|
| | ≥ 2 CKD Diagnosis Codes [b] N = 299,751 (49.1%) | 1 CKD Diagnosis Code and ≥ 1 low eGFR N = 50,540 (8.3%) | ≥ 2 Low eGFRs (no diagnosis) N = 77,855 (12.8%) | 1 CKD Diagnosis only N = 31,948 (5.2%) | 1 low eGFR only N = 150,158 (24.6 %) | Total N = 610,252 |
| Institutional stay during baseline, N (%) yes | 14,715 (4.9) | 2,348 (4.6) | 2,728 (3.5) | 2,026 (6.3) | 5,527 (3.7) | 27,344 (4.5) |
| Comorbidity score, [69] mean (SD) [e] | 1.9 (2.5) | 1.6 (2.5) | 0.8 (1.9) | 1.7 (2.4) | 1.0 (2.1) | 1.5 (2.4) |
| Selected individual comorbidities | | | | | | |
| Congestive heart failure | 75,417 (25.2) | 10,071 (19.9) | 10,650 (13.7) | 6,665 (20.9) | 18,819 (12.5) | 121,622 (19.9) |
| HIV/AIDS | 727 (0.2) | 139 (0.3) | 157 (0.2) | 65 (0.2) | 391 (0.3) | 1,479 (0.2) |
| Hypertension | 260,679 (87.0) | 42,490 (84.1) | 65,424 (84.0) | 26,091 (81.7) | 111,818 (74.5) | 506,502 (83.0) |
| Pulmonary disease, chronic | 79,935 (26.7) | 12,536 (24.8) | 14,620 (18.8) | 8,613 (27.0) | 29,998 (20.0) | 145,702 (23.9) |
| Peripheral vascular disease | 85,222 (28.4) | 11,508 (22.8) | 11,420 (14.7) | 7,661 (24.0) | 21,745 (14.5) | 137,556 (22.5) |
| Tumor, any | 41,298 (13.8) | 7,250 (14.3) | 9,099 (11.7) | 4,335 (13.6) | 18,587 (12.4) | 80,569 (13.2) |
| Diabetes, any | 148,949 (49.7) | 19,863 (39.3) | 28,954 (37.2) | 14,537 (45.5) | 43,212 (28.8) | 255,515 (41.9)) |
| Myocardial infarction or stroke | 42,136 (14.1) | 6,489 (12.8) | 7,952 (10.2) | 4,453 (13.9) | 14,322 (9.5) | 75,352 (12.3) |
| Previous CKD diagnosis in 2011 | 209,807 (70.0) | 15,185 (30.0) | 8,633 (11.1) | 11,953 (37.4) | 10,368 (6.9) | 255,946 (41.9) |
| Mean follow-up time (SD) [f] | 340.4 (73.7) | 336.9 (80.9) | 359.6 (27.4) | 311.9 (111.2) | 333.6 (86.1) | 339.4 (76.7) |
| Death within one year | 17,847 (6.0) | 3,874 (7.7) | 878 (1.1) | 2,260 (7.1) | 5,968 (4.0) | 30,827 (5.1) |

[a] Covariates assessed for the 183 days prior to cohort entry date (T0 or T0L) except as noted
[b] Patients with ≥ 2 CKD diagnosis codes within 365 days were assigned to this group whether or not they also had low eGFRs available. Many also had eGFRs available.
[c] Based on second coded diagnosis for those with 2 coded diagnoses
[d] Serum creatinine to calculate low eGFR not available in 2012 for entire 100% because 365 day follow-up could extend into 2013
[e] Determined over the 183 days prior to the cohort entry date. Diagnosis of renal failure is one of the coded conditions usually included in calculating this comorbidity score. However, because this work was intended to identify a cohort of individuals with CKD, coded diagnosis of renal failure was excluded from the comorbidity score calculated for this cohort.
[f] Maximum follow-up time assessed was 365 days

The mean (SD) age of patients in the overall CKD cohort was 73.1 (9.9) years. This varied from 70.4 years for patients with only one CKD diagnosis to 74.4 years for patients with two or more low eGFR and no coded CKD diagnosis. For all subgroups except those in the "1 CKD diagnosis only" subgroup (5.2% of the cohort), the proportion of patients identified with CKD was higher in the oldest age group (75 – 89 years) than in the younger age categories (< 65 or 65 – 74 year; Table 30). Overall, 55.9% of the total cohort was female, ranging from 49.1% of those with only one CKD diagnosis to 63.8% of those with at least two low eGFR measurements.

As shown by the total comorbidity score, individuals who entered the cohort through at least 2 low eGFR had less comorbidity (mean 0.8) than the cohort overall (mean 1.5), particularly relative to those

who had at least two coded CKD diagnoses (mean 1.9). Hypertension (83.0%) and diabetes (41.9%) were the most common comorbidities (Table 30).

Across the three cohort subgroups that had any coded CKD diagnosis (N = 382,239), 60.8% were coded as Stage 3, 4.6% were coded as Stage 4, 0.4% were coded as Stage 5, and 34.3% had a coded diagnosis that was not staged. However, when CKD stage was estimated for the entire cohort using available eGFR or coded CKD diagnosis (N = 610,252), 84.0% were Stage 3, 6.6% were Stage 4, 0.9% were Stage 5, and only 8.5% remained not staged. When Stage was estimated for the entire cohort, site 3 had the lowest proportion of Stage 3 (81.0%), while site 1 had the highest proportion (91.9%). Conversely, site 3 had the highest proportion with no stage assignment (11.4%)(Appendix H).

Although 70% (N = 209,870) of the cohort with two or more CKD diagnoses in 2012 also had a diagnosis of CKD in 2011 (Table 30), that sub-cohort accounted for 82% of the cohort with a prior CKD diagnosis (209,807 of 255,946). In contrast, only 11.1% (N = 8633) of the cohort identified by ≥ 2 low eGFRs (no CKD coded diagnosis) in 2012 had a prior CKD diagnosis, accounting for only 3.4% of those with a prior diagnosis (8633 of 255,946).

### c. Cohort Identification Test Case 2 results and discussion

Many database studies of patients with CKD require two indicators of CKD (e.g., qualifying laboratory result and/or coded diagnosis) for cohort inclusion. Considering only patients with at least 2 CKD diagnoses in our cohort identified 299,751 individuals. As shown in Table 31, augmenting this cohort with patients who had one coded diagnosis plus at least one low eGFR (N = 50,540) and with patients who had at least two low eGFR (N = 77,855) increased the cohort to 428,146 individuals, a 30.0% increase in cohort size over including only patients with at least two CKD diagnoses. The overwhelming majority of these additional individuals were Stage 3 (76,546 of 77,855). Site-specific cohort augmentation from estimated eGFR was 28.7% at site 2, 30.0% at site 3, and 36.2% at site 1 (Appendix H).

**Table 31. Identification of the 2012 Cohort of Patients with Chronic Kidney Disease using an Electronic Data Definition that Requires at Least Two Coded Diagnoses: Augmentation of the Cohort Identified Using Laboratory Test Results Criteria**

| | Patients Identified Using Coded Diagnosis Definition Requiring ≥ 2 Diagnoses (with or without eGFR < 60 ml/min/1.73m²) | Additional Patients Identified Using Laboratory Test Results (No or 1 Diagnosis with eGFR < 60 ml/min/1.73m²) | | Total Patients in CKD Cohort |
|---|---|---|---|---|
| | | Patients with 1 Coded Diagnosis and ≥ 1 eGFR < 60 ml/min/1.73m² | Patients with ≥ 2 eGFR < 60 ml/min/1.73m² (No Coded Diagnosis) | |
| N (%) in Subgroup | 299,751 (70.0) [a] | 50,540 (11.8) | 77,855 (18.2) | 428,146 |
| Subtotal | | 128,395 (30.0) | | |

[a] Comprised of 85,436 (20.0%) with ≥ 2 coded diagnoses only and 214,315 (50.1%) with ≥ 2 coded diagnoses and ≥ 1 eGFR < 60 ml/min/1.73m²

As shown in Table 32, the 2008 and 2010 cohorts increased by greater proportions (38.6% and 31.5%, respectively) than the 2012 cohort (30.0%) when patients with one coded diagnosis plus at least one low eGFR and patients with at least two low eGFR were included. The contribution to the cohort proportion by including patients with one coded diagnosis and at least one eGFR was relatively consistent across the three cohort years both across and within each of the three sites. However, the contribution to the cohort proportion by including patients with two or more eGFR contributed 26.2% in 2008, declining to 18.2% in 2012. The shift was dramatic at sites 1 and 2 where the additional percentages of individuals contributed to the cohort in 2008 and 2012 respectively were 34.5% and 21.7% (site 1) and 27.8% and 15.0% (site 2). This could suggest not only increased prevalence of CKD, but also expanded recognition of low eGFR and improved assignment of CKD diagnosis over time.

**Table 32. Augmentation by Year of the Identified Chronic Kidney Disease Cohort Using Laboratory Test Results Criteria in addition to at Least Two Coded Diagnoses of Chronic Kidney Disease**

| Cohort Year | Proportion of Patients Identified Using Coded Diagnosis Definition Requiring ≥ 2 Diagnoses (with or without eGFR < 60 ml/min/1.73m²) | Proportions of Patients Identified Using Laboratory Test Results (No or 1 Diagnosis with eGFR < 60 ml/min/1.73m²) | | Percent Total Increase in Cohort Size | Total Patients in CKD Cohort |
|---|---|---|---|---|---|
| | | Patients with 1 Coded Diagnosis and ≥ 1 eGFR < 60 ml/min/1.73m² | Patients with ≥ 2 eGFR < 60 ml/min/1.73m² (No Coded Diagnosis) | | |
| All Sites | | | | | |
| 2012 | 70.0 | 11.8 | 18.2 | 30.0 | 428,146 |
| 2010 | 68.6 | 11.0 | 20.5 | 31.5 | 341,263 |
| 2008 | 61.4 | 12.4 | 26.2 | 38.6 | 251,409 |
| Site 1 | | | | | |
| 2012 | 63.8 | 14.4 | 21.7 | 36.1 | 27,560 |
| 2010 | 60.9 | 13.2 | 25.9 | 39.1 | 22,084 |
| 2008 | 50.7 | 14.8 | 34.5 | 48.3 | 17,638 |
| Site 2 | | | | | |
| 2012 | 71.3 | 13.7 | 15.0 | 28.7 | 130,900 |
| 2010 | 65.5 | 12.7 | 21.7 | 34.4 | 112,733 |
| 2008 | 57.7 | 14.5 | 27.8 | 42.3 | 100,152 |
| Site 3 | | | | | |
| 2012 | 70.0 | 10.6 | 19.4 | 30.0 | 269,686 |
| 2010 | 71.1 | 9.7 | 19.2 | 28.9 | 206,446 |
| 2008 | 65.5 | 10.5 | 23.9 | 34.4 | 133,619 |

As shown in **Table 33**, if the less stringent criterion of one or more coded CKD diagnosis is used, adding patients identified by eGFR calculated using creatinine values augments the 2012 CKD cohort proportion from the participating sites by 16.9%. Site-specific cohort augmentation was 14.4% at site 2, 17.7% at site 3, and 20.9% at site 1 (Appendix H).

**Table 33. Identification of the 2012 Cohort of Patients with Chronic Kidney Disease using an Electronic Data Definition that requires at Least One Coded Diagnosis: Augmentation of the Cohort Identified Using Laboratory Test Results Criteria**

| | Patients Identified Using Coded Diagnosis Definition Requiring ≥ 1 Diagnoses (with or without eGFR < 60 ml/min/1.73m²) | Lab Results ≥ 2 eGFR < 60 ml/min/1.73m² | Total Patients in CKD Cohort |
|---|---|---|---|
| N (%) in Subgroup | 382,239 (83.1) [a] | 77,855 (16.9) | 460,094 |

[a] Comprised of 31,948 (6.9%) with 1 coded diagnosis only, 50,540 (11.0%) with 1 coded diagnosis and ≥ 1 eGFR < 60 ml/min/1.73m², 85,436 (18.6%) with ≥ 2 coded diagnoses only, and 214,315 (46.6%) with >=2 coded diagnoses and ≥ 1 eGFR < 60 ml/min/1.73m²

As shown in Table 34, even when the less stringent criterion of one or more coded CKD diagnosis is used as the initial criterion, the 2008 and 2010 cohorts increased by greater proportions (24.3% and 19.0%, respectively) than the 2012 cohort (16.9%) when the criterion of including patients in the cohort who only had two or more low eGFR and no CKD diagnosis was added. The cohort additions were again greater at sites 1 and 2 where the additional proportions of individuals in the cohort in 2008 and 2012 respectively were 33.3% and 20.9% (site 1) and 27.1% and 14.4% (site 2).

**Table 34. Augmentation by Year of the Identified Chronic Kidney Disease Cohort Using Laboratory Results Criteria in addition to at Least One Coded Diagnosis of Chronic Kidney Disease**

| Cohort Year | Patients Identified Using Coded Diagnosis Definition Requiring ≥ 1 Diagnoses (with or without eGFR < 60 ml/min/1.73m²), % of Total | Lab Results ≥ 2 eGFR < 60 ml/min/1.73m², % of Total | Total Patients in CKD Cohort |
|---|---|---|---|
| All Sites | | | |
| 2012 | 83.1 | 16.9 | 460,094 |
| 2010 | 81.0 | 19.0 | 366,842 |
| 2008 | 75.7 | 24.3 | 271,096 |
| Site 1 | | | |
| 2012 | 79.1 | 20.9 | 28,665 |
| 2010 | 75.1 | 24.9 | 22,996 |
| 2008 | 66.7 | 33.3 | 18,279 |
| Site 2 | | | |
| 2012 | 85.6 | 14.4 | 136,080 |
| 2010 | 78.9 | 21.1 | 116,164 |
| 2008 | 72.9 | 27.1 | 102,824 |
| Site 3 | | | |
| 2012 | 82.3 | 17.7 | 295,349 |
| 2010 | 82.6 | 17.4 | 227,682 |
| 2008 | 78.7 | 21.3 | 149,993 |

In Tables 32 through 34, the cohort includes adults with the first indicator of CKD in calendar year 2012, 2010, or 2008. Additional diagnoses or low eGFR results were captured only up to 365 days after that first indicator.

***Percent agreement between pairs of CKD indicators***

Percent agreements between CKD indicators in the 2012 cohort are shown in Table 35. The first percent agreement shown is between CKD identified from at least 1 coded diagnosis (N = 383,239) compared to identifying CKD using at least 2 eGFR values <60ml/min/1.73m$^2$ measured at least 90 days apart (N = 246,135). Overall, there was agreement between these diagnosis and laboratory result indicators for 52.2% of the 610,252 patients, including 27.6% that had both indicators and 24.6% that had neither indicator. Site variation was evident, ranging from 47.2% agreement at site 3 to 63.1% agreement at site 1.

**Table 35. Agreement between Different Definitions of Chronic Kidney Disease in a Cohort of Adults with at least one Chronic Kidney Disease Diagnosis or at least one eGFR < 60 ml/min/1.73m2 in 2012**

| Percent Agreement between Chronic Kidney Disease Diagnosis and eGFR Indicators | Total | Site 1 | Site 2 | Site 3 |
|---|---|---|---|---|
| Total in 2012 cohort, N | 610,252 | 46,886 | 185,564 | 377,802 |
| | | | | |
| ≥ 1 CKD diagnosis (N = 383,239) and ≥ 2 eGFR < 60 (N = 246,135) % agreement | 52.2 | 63.1 | 59.5 | 47.2 |
| % that agree because both indicators are present | 27.6 | 24.2 | 32.9 | 25.4 |
| % that agree because neither indicator is present | 24.6 | 38.9 | 26.7 | 21.8 |
| | | | | |
| ≥ 2 CKD diagnosis (N = 299,751) and ≥ 2 eGFR < 60 (N = 246,135) % agreement | 58.0 | 67.0 | 63.5 | 54.3 |
| % that agree because both indicators are present | 23.8 | 20.8 | 28.6 | 21.7 |
| % that agree because neither indicator is present | 34.3 | 46.3 | 34.9 | 32.5 |
| | | | | |
| > 1 CKD diagnosis (N = 383,239) and > 1 eGFR < 60 (N = 498,666) % agreement | 44.4 | 41.5 | 53.3 | 40.3 |
| % that agree because both indicators are present | 44.4 | 41.5 | 53.3 | 40.3 |
| % that agree because neither indicator is present [a] | 0 | 0 | 0 | 0 |
| [a] Cohort selection required at least one CKD diagnosis or one eGFR < 60 ml/min/1.73m$^2$ | | | | |

The percent agreement between CKD identified using at least 2 coded diagnoses (N = 299,751) compared to using at least 2 eGFR values <60ml/min/1.73m$^2$ measured at least 90 days apart is also shown in Table 35. Among the CKD definitions compared in Table 35, these two indicators had the highest percent agreement at 58.0% (23.8% meeting both indicators and 34.3% meeting neither indicator). Site variation was evident, with the percent agreement ranging from a low of 54.3% at site 3 to a high of 67.0% at site 1.

Finally, the percent agreement between CKD identified from at least 1 coded diagnosis compared to at least 1 eGFR value <60ml/min/1.73m$^2$ (N = 498,666) is shown in Table 35. Across sites, these two indicators had the lowest percent agreement with 44.4% having both indicators. Because having at least one CKD diagnosis or at least one eGFR < 60 ml/min/1.73m$^2$ was required for cohort inclusion, 0% had neither indicator. Percent agreement varied from 40.3% at site 3 to 53.3% at site 2.

### d. Summary of Cohort Identification Test Case 2

The number of adults identified for inclusion in a CKD cohort was substantially augmented by estimating GFR using available serum creatinine results in addition to coded diagnosis data. This finding was relatively consistent across sites and years. It was also true regardless of whether more or less stringent CKD indicators were applied for cohort inclusion, with the proportion of the 2012 CKD cohort identified using laboratory results varying from 16.9% to 30.0%. Similar patterns were observed in the 2008 and 2010 CKD cohorts, but these earlier cohorts increased by even greater proportions when individuals identified as having CKD using eGFR data were considered (38.6% and 31.5%, respectively). That a lower proportion of individuals were identified only from eGFR data in the most recent cohort is potentially consistent with improved assignment of CKD coded diagnosis over time.

Patients identified only through two or more eGFR results had less comorbidity than those identified through two or more CKD coded diagnoses. Stage 3 CKD was by far the most common CKD stage, whether determined from a CKD diagnosis code or from estimated eGFR.

Overall, 55.9% of the total cohort was female; however 63.8% of individuals identified with at least two low eGFR measurements but no diagnosis was female. This raises questions about why more women were identified from eGFR alone. For example, do more women have creatinine values in the "normal" range despite low eGFR, making clinicians less likely to consider CKD? This would be consistent with published evidence that women are underdiagnoses with CKD relative to men, particularly in practices where laboratory results are reported as serum creatinine without eGFR.[97] Alternatively, do women have more healthcare visits than men and this finding is aligned with women having more opportunities to have laboratory monitoring conducted? Potential clinical and safety outcomes, policy, and research implications of this observation will be important to explore in future work.

The most stringent CKD definitions had the highest percent agreement. The CKD indicator set with the higher proportion having both indicators present was the least stringent, but the least restrictive CKD definition also had the lowest percent agreement. This finding is important for future CKD cohort construction because it implies that requiring more stringent criteria improves specificity. The requirement for this cohort that both indicators must have been present within the 365 day period likely resulted in percent agreements lower than would have been found had a longer time horizon been used. We recommend this be assessed in a future project.

Strengths of this test case include the cohort size, examining data from three disparate healthcare delivery systems, and applying several different indicator sets to identify the CKD cohorts. A limitation of this test case was not considering urine protein laboratory test results in defining patients with CKD. The MSDD does not include urine protein results so these laboratory test results were not available for use.

## 3. Summary of the performance of laboratory test results in augmenting cohort identification

These analyses estimated whether additional individuals were included in two distinct cohort types when relevant laboratory results were considered in addition to coded diagnoses and procedures. The cohorts were a cohort of pregnant women and a cohort of patients with CKD. The laboratory tests were HCG results and serum creatinine results (used to determine eGFR using the CKD-EPI equation).

In both test cases, the cohort size was augmented through inclusion of the laboratory results. In the pregnant cohort, the number of women identified as pregnant during the prenatal period solely due to positive HCG results was 5.1%. However, the proportion of pregnancies identified by only an HCG result

varied dramatically across pregnancy outcomes, from 0% of live births to 32.6% of pregnancies that ended in abortion. Fully, 52.4% of the cohort had at least one HCG result. Using positive HCG results enabled cohort inclusion 11.6 to 15.7 days earlier than using only prenatal diagnosis or procedure codes. **The vast majority of positive HCG results fell within the 270 day gestational age metric, suggesting that continued use of this metric in database studies in determining gestational age from live birth outcomes is appropriate. In future work it will be important to study the subgroup of women who only had a laboratory-based indicator of pregnancy in the observational database, in particular within the context of implications related to risks of drug exposures during pregnancy.**

The size of the cohort of adults with CKD was substantially augmented by estimating GFR using creatinine laboratory results in addition to coded diagnosis. The most stringent CKD definitions had the highest percent agreement. Our cohort was only followed for 365 days -- it is likely that these percent agreements are lower than would have been found had a longer time horizon been used. **We recommend that serum creatinine results available in the MSDD be included to supplement cohort identification criteria when constructing a CKD cohort for medical product safety surveillance activities.**

## F. OUTCOME DETECTION TEST CASES

The purpose of the Outcomes Detection Test Cases was to examine methods for using laboratory results to improve claims-based definitions of health outcomes, including determining the availability and nature of missingness of laboratory results and whether and how strategies might need to differ by type of laboratory result. In addition, in these test cases, the Workgroup determined whether additional individuals with a health outcome of interest were identified when available laboratory results were considered as criteria for the outcome definition in addition to coded diagnoses, procedures, and medications. The characteristics of individuals identified using outcomes criteria that did and did not include laboratory results values are described, the proportions of the cohorts with laboratory results outcomes data available are provided, and cohort-specific research questions are examined.

### 1. Outcomes Detection Test Case 1: Type 2 diabetes mellitus (T2DM) and blood glucose or glycosylated hemoglobin laboratory test results values outcomes among adults initiating a Second Generation Antipsychotic Agent

#### a. Outcomes Detection Test Case 1 cohort development

As discussed in Baseline Confounder Test Case 1, SGA are prescribed to aid in treatment of mental health disorders and carry metabolic risks such as development of T2DM.[64, 65] The intent of Outcomes Detection Test Case 1 was to determine whether including GLU results (fasting glucose, random glucose, or HbA1c) obtained within 365 days after initiation of a SGA ("follow-up GLU") increased the number of individuals preliminarily identified as developing T2DM.

Research questions addressed include:

1. Does inclusion of follow-up GLU result value(s) after SGA initiation identify additional individuals with preliminary indicators of T2DM?
    a. How many additional outcomes are identified?
    b. How many cases of T2DM are identified earlier also using follow-up GLU results versus only diagnoses codes and antidiabetes medication dispensings?

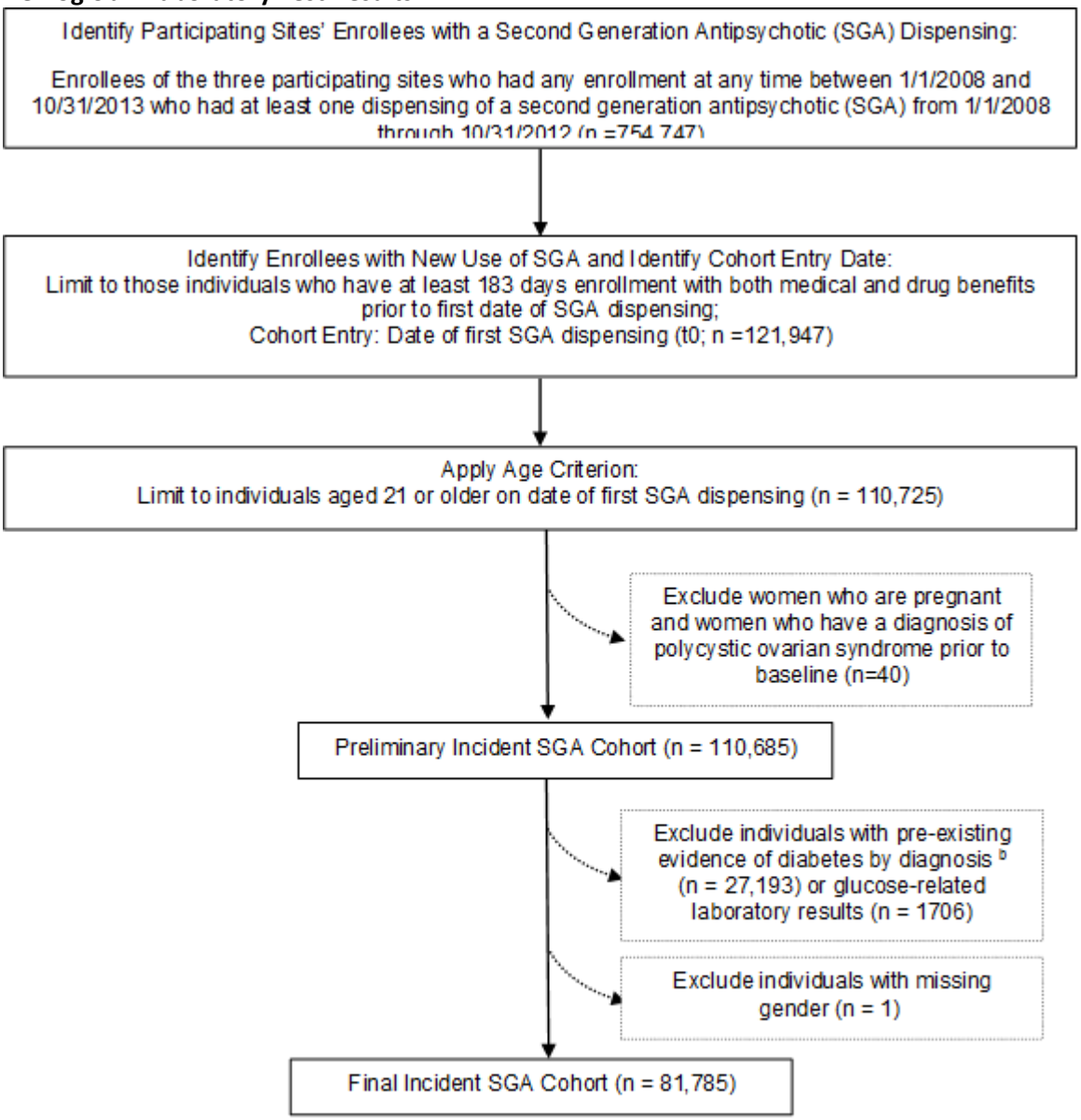
2. Regarding the timing after SGA initiation until the observed follow-up GLU value:
    a. What is the distribution of time from the cohort entry date to the observed GLU result values?
    b. Does the time from the cohort entry date to available GLU result value vary by SGA?
3. When is the appropriate time to censor patients from the cohort if no follow-up GLU result value is observed?
4. What are the considerations around imputing GLU outcomes?

The workgroup also addressed the following clinical question: Does risk of developing T2DM within 365 days after SGA initiation differ by specific SGA?

Exposures included SGA newly-started between July 1, 2008 and October 31, 2012 in adults $\geq$ 21 years of age who had a minimum of 183 days enrollment in the health plan with medical and drug coverage prior to the first dispensing of an SGA and who did not have a history of diabetes (defined as no dispensing of an antidiabetic medication, no coded diabetes diagnosis, and no elevated GLU result value within the 183 days prior to SGA initiation). The cohort entry date ($T_0$) was the date of the first SGA dispensing. The following SGA (single and combination products) were included to identify the preliminary cohort initiating an SGA: aripiprazole, asenapine, iloperidone, lurasidone, olanzapine, paliperidone, quetiapine, risperidone, and ziprasidone. Because use of several SGAs was very low, the final cohort included only individuals exposed to aripiprazole, olanzapine, quetiapine, or risperidone. We excluded pregnant women and women with polycystic ovarian syndrome. Details of cohort development are shown in Figure 7. Individuals in the cohort were censored at the first of any of the following: death, discontinuation of medical or drug coverage, diabetes outcome, or end of study period. Covariates considered for this test case are shown in Appendix D.

Diabetes outcome was defined as the presence of any single inpatient or outpatient diabetes diagnosis code (ICD-9 CM 250.x), antidiabetic medication dispensing, or GLU result value compatible with diabetes (fasting glucose $\geq$ 126mg/dl), random glucose $\geq$ 200mg/dl, or HbA1c $\geq$ 6.5%) within 365 days after $T_0$. Any of these diabetes indicators were considered a "preliminary indication of diabetes." This set of diabetes indicators provided a sensitive (less specific) definition of diabetes. The workgroup considered these outcomes preliminary rather than confirmatory indicators of diabetes, and considered that, for the purposes of this test case, a preliminary indication of diabetes was applicable.

**Figure 7. Test Case for Outcomes Detection Test Case 1, Diabetes Outcome among Adults after Initiation of a Second Generation Antipsychotic Agent with Missing Blood Glucose or Glycosylated Hemoglobin Laboratory Test Results [a]**




[a] Definitions and codes are in Appendix D

[b] Any of the following within the 183 days prior to cohort entry: Inpatient or outpatient diabetes diagnosis, dispensing of anti-diabetic medication, hemoglobin $A_{1c} \geq 6.5\%$; fasting plasma glucose

### b. Outcome Detection Test Case 1 descriptive analysis

Descriptions of the cohort by SGA initiated and by site are in Tables 36 and 37.The most common SGA started was quetiapine (48%), followed by risperidone (24%) and aripiprazole (18%), with olanzapine being used least frequently (11%) of these SGAs. SGA use varied by site. For example, site 2 contributed 37.5% of the cohort, but 41.8% of quetiapine users, while site 3 contributed 57.9% of the cohort and 62.8% of risperidone users. The cohort averaged 60.3 years of age; 62.3% were female.

During the year following SGA initiation, 6.1% of the cohort developed a preliminary indication of diabetes, varying from 5.2% of individuals' dispensed aripiprazole to 6.5% of individuals' dispensed olanzapine or risperidone. Preliminary indicators of diabetes varied across sites (Table 37) from 3.7% and 3.8% at sites 1 and 2 respectively, to 7.8% at site 3.

**Table 36. Characteristics of Individuals in the Outcomes Detection Test Case 1 Population, Initiation of a Second Generation Antipsychotic and Type 2 Diabetes Mellitus across Three Sites, by Specific Second Generation Antipsychotic Agent**

| Variable | Second Generation Antipsychotic Agent | | | | |
|---|---|---|---|---|---|
| | Aripiprazole (N = 14,333) | Olanzapine (N = 8,864) | Quetiapine (N = 39,353) | Risperidone (N = 19,235) | Overall (N = 81,785) |
| Diabetes outcome, N (%) Yes [a] | 744 (5.2) | 574 (6.5) | 2,411 (6.1) | 1,259 (6.5) | 4,988 (6.1) |
| Any Baseline GLU, N (%) Yes | 4,789 (33.4) | 3,770 (42.5) | 16,813 (42.7) | 7,745 (40.3) | 33,117 (40.5) |
| Baseline HbA1c, N (%) Yes | 681 (4.8) | 437 (4.9) | 1,927 (4.9) | 946 (4.9) | 3,991 (4.9) |
| Baseline Fasting Glucose, N (%) Yes | 1,436 (10.0) | 825 (9.3) | 3,724 (9.5) | 1,526 (7.9) | 7,511 (9.2) |
| Baseline Random Glucose, N (%) Yes | 3,615 (25.2) | 3,319 (37.4) | 14,452 (36.7) | 6,724 (35.0) | 28,110 (34.4) |
| Any follow-up GLU, N (%) Yes | 6,415 (44.8) | 4,030 (45.5) | 18,711 (47.5) | 8,802 (45.8) | 37,958 (46.4) |
| Any follow-up HbA1c, N (%) Yes | 1,361 (9.5) | 741 (8.4) | 3,290 (8.4) | 1,643 (8.5) | 7,035 (8.6) |
| Any follow-up fasting glucose, N (%) Yes | 2,547 (17.8) | 1,382 (15.6) | 5,808 (14.8) | 2,820 (14.7) | 12,557 (15.4) |
| Any follow-up random glucose, N (%) Yes | 4,561 (31.8) | 3,192 (36.0) | 15,149 (38.5) | 6,988 (36.3) | 29,890 (36.5) |
| Gender, N (%) female | 10,028 (70.0) | 5,172 (58.3) | 23,935 (60.8) | 11,786 (61.3) | 50,921 (62.3) |
| Age at cohort entry, mean (SD) years | 50.1 (15.7) | 61.0 (20.3) | 62.0 (20.5) | 64.2 (21.0) | 60.3 (20.4) |
| Year of cohort entry, N (%) | | | | | |
| 2008 | 1,783 (12.4) | 1,345 (15.2) | 5,091 (12.9) | 2,396 (12.5) | 10,615 (13.0) |
| 2009 | 3,826 (26.7) | 2,046 (23.1) | 8,941 (22.7) | 4,350 (22.6) | 19,163 (23.4) |
| 2010 | 3,361 (23.4) | 1,839 (20.7) | 8,730 (22.2) | 4,345 (22.6) | 18,275 (22.3) |
| 2011 | 2,774 (19.4) | 1,846 (20.8) | 8,750 (22.2) | 4,309 (22.4) | 17,679 (21.6) |
| 2012 | 2,589 (18.1) | 1,788 (20.2) | 7,841 (19.9) | 3,835 (19.9) | 16,053 (19.6) |
| Site, N (%) | | | | | |
| 1 | 448 (3.1) | 220 (2.5) | 1,525 (3.9) | 1,571 (8.2) | 3,764 (4.6) |
| 2 | 5,208 (36.3) | 3,405 (38.4) | 16,449 (41.8) | 5,575 (29.0) | 30,637 (37.5) |
| 3 | 8,677 (60.5) | 5,239 (59.1) | 21,379 (54.3) | 12,089 (62.8) | 47,384 (57.9) |
| Hispanic, N (%) | 715 (5.0) | 511 (5.8) | 2,320 (5.9) | 1,132 (5.9) | 4,678 (5.7) |
| Race, N (%) | | | | | |
| White | 9,095 (63.5) | 6,217 (70.1) | 29,060 (73.8) | 14,097 (73.3) | 58,469 (71.5) |
| African American | 684 (4.8) | 653 (7.4) | 2,867 (7.3) | 1,909 (9.9) | 6,113 (7.5) |
| Other | 454 (3.2) | 459 (5.2) | 1,355 (3.4) | 754 (3.9) | 3,022 (3.7) |
| Unknown | 4,100 (28.6) | 1,535 (17.3) | 6,071 (15.4) | 2,475 (12.9) | 14,181 (17.3) |
| GLU CPT code during baseline, N (%) yes [b] | 4,270 (29.8) | 4,418 (49.8) | 16,986 (43.2) | 8,164 (42.4) | 33,838 (41.4) |
| GLU CPT code during follow-up, N (%) yes [b] | 4,915 (34.3) | 3,979 (44.9) | 17,429 (44.3) | 8,151 (42.4) | 34,474 (42.2) |
| Number of unique medication classes dispensed during baseline, mean (SD) | 10.8 (4.8) | 9.7 (5.2) | 10.1 (4.9) | 9.5 (4.8) | 10.0 (4.9) |

| Variable | Second Generation Antipsychotic Agent | | | | |
|---|---|---|---|---|---|
| | Aripiprazole (N = 14,333) | Olanzapine (N = 8,864) | Quetiapine (N = 39,353) | Risperidone (N = 19,235) | Overall (N = 81,785) |
| Number of ambulatory healthcare visits during baseline, mean (SD) | 13.0 (15.3) | 14.9 (18.3) | 14.7 (19.2) | 13.6 (16.3) | 14.2 (17.8) |
| Any emergency department visit during baseline, N (%) yes | 3,728 (26.0) | 4,043 (45.6) | 15,745 (40.0) | 7,525 (39.1) | 31,041 (38.0) |
| Any hospitalization during baseline, N (%) yes | 1,835 (12.8) | 3,169 (35.8) | 10,864 (27.6) | 5,558 (28.9) | 21,426 (26.2) |
| Any institutional stay during baseline, N (%) yes | 2,003 (14.0) | 2,479 (28.0) | 9,063 (23.0) | 5,026 (26.1) | 18,571 (22.7) |
| Comorbidity score, mean (SD) [c, d] | 1.2 (1.5) | 2.3 (2.6) | 2.0 (2.4) | 2.2 (2.4) | 1.9 (2.3) |
| Individual comorbidities, N (%) yes [c, d] | | | | | |
| Alcohol abuse | 894 (6.2) | 741 (8.4) | 3,566 (9.1) | 1,312 (6.8) | 6,513 (8.0) |
| Anemia | 1,221 (8.5) | 1,536 (17.3) | 6,113 (15.5) | 3,206 (16.7) | 12,076 (14.8) |
| Cardiac arrhythmia | 796 (5.6) | 1,405 (15.9) | 5,856 (14.9) | 3,034 (15.8) | 11,091 (13.6) |
| Coagulopathy | 165 (1.2) | 305 (3.4) | 1,172 (3.0) | 559 (2.9) | 2,201 (2.7) |
| Heart failure, chronic | 522 (3.6) | 1,054 (11.9) | 4,328 (11.0) | 2,379 (12.4) | 8,283 (10.1) |
| Dementia | 441 (3.1) | 1,604 (18.1) | 7,777 (19.8) | 5,047 (26.2) | 14,869 (18.2) |
| Fluid/electrolyte disorder | 933 (6.5) | 1,776 (20.0) | 6,305 (16.0) | 3,199 (16.6) | 12,213 (14.9) |
| HIV | 68 (0.5) | 36 (0.4) | 171 (0.4) | 49 (0.3) | 324 (0.4) |
| Hypertension | 4,333 (30.2) | 4,147 (46.8) | 18,615 (47.3) | 9,714 (50.5) | 36,809 (45.0) |
| Hemiplegia | 123 (0.9) | 154 (1.7) | 718 (1.8) | 321 (1.7) | 1,316 (1.6) |
| Liver disease | 326 (2.3) | 258 (2.9) | 1,090 (2.8) | 430 (2.2) | 2,104 (2.6) |
| Metastatic cancer | 77 (0.5) | 291 (3.3) | 614 (1.6) | 224 (1.2) | 1,206 (1.5) |
| Psychosis | 9,918 (69.2) | 5,227 (59.0) | 19,447 (49.4) | 11,199 (58.2) | 45,791 (56.0) |
| Pulmonary disease | 2,142 (14.9) | 1,845 (20.8) | 7,343 (18.7) | 3,800 (19.8) | 15,130 (18.5) |
| Pulmonary circulation disorder | 90 (0.6) | 171 (1.9) | 753 (1.9) | 408 (2.1) | 1,422 (1.7) |
| PVD | 537 (3.7) | 910 (10.3) | 3,893 (9.9) | 2,210 (11.5) | 7,550 (9.2) |
| Renal | 444 (3.1) | 832 (9.4) | 3,820 (9.7) | 2,271 (11.8) | 7,367 (9.0) |
| Tumor | 550 (3.8) | 821 (9.3) | 2,799 (7.1) | 1,348 (7.0) | 5,518 (6.7) |
| Weight loss | 132 (0.9) | 452 (5.1) | 1,291 (3.3) | 588 (3.1) | 2,463 (3.0) |
| Individual comorbidities specific to test case [c] | | | | | |
| Myocardial infarction, acute | 273 (1.9) | 412 (4.6) | 1,942 (4.9) | 988 (5.1) | 3,615 (4.4) |
| Ischemic Stroke | 407 (2.8) | 672 (7.6) | 3,006 (7.6) | 1,679 (8.7) | 5,764 (7.0) |
| Intracranial Hemorrhage | 49 (0.3) | 128 (1.4) | 596 (1.5) | 234 (1.2) | 1,007 (1.2) |
| Osteoarthritis | 1,744 (12.2) | 1,582 (17.8) | 7,230 (18.4) | 3,602 (18.7) | 14,158 (17.3) |
| Depression | 9,952 (69.4) | 4,589 (51.8) | 20,003 (50.8) | 9,313 (48.4) | 43,857 (53.6) |

[a] Within 365 days after the cohort entry date
[b] CPT codes 80047, 80048, 80050, 80053, 80069, 82947, 83036, 83037
[c] Determined over the 183 days prior to the cohort entry date
[d] Gagne et al [69]

**Table 37. Characteristics of Individuals in the Outcomes Detection Test Case 1 Population, Initiation of a Second Generation Antipsychotic and Type 2 Diabetes Mellitus, by Site**

| Variable | Data Partner Site | | | |
|---|---|---|---|---|
| | Site 1 (N = 3,764) | Site 2 (N = 30,637) | Site 3 (N = 47,384) | All Sites (N = 81,785) |
| Diabetes outcome, N (%) Yes [a] | 141 (3.7) | 1,164 (3.8) | 3,683 (7.8) | 4,988 (6.1) |
| Any Baseline GLU, N (%) Yes | 2,154 (57.2) | 17,973 (58.7) | 12,990 (27.4) | 33,117 (40.5) |
| Baseline HbA1c, N (%) Yes | 105 (2.8) | 2,170 (7.1) | 1,716 (3.6) | 3,991 (4.9) |
| Baseline Fasting Glucose, N (%) Yes | 622 (16.5) | 6,738 (22.0) | 151 (0.3) | 7,511 (9.2) |
| Baseline Random Glucose, N (%) Yes | 1,790 (47.6) | 13,575 (44.3) | 12,745 (26.9) | 28,110 (34.4) |
| Any follow-up GLU, N (%) Yes | 2,567 (68.2) | 20,268 (66.2) | 15,123 (31.9) | 37,958 (46.4) |
| Any follow-up HbA1c, N (%) Yes | 231 (6.1) | 3,627 (11.8) | 3,177 (6.7) | 7,035 (8.6) |
| Any follow-up fasting glucose, N (%) Yes | 1,196 (31.8) | 10,991 (35.9) | 370 (0.8) | 12,557 (15.4) |
| Any follow-up random glucose, N (%) Yes | 1,972 (52.4) | 13,229 (43.2) | 14,689 (31.0) | 29,890 (36.5) |
| Gender, N (%) female | 2,402 (63.8) | 19,103 (62.4) | 29,416 (62.1) | 50,921 (62.3) |
| Age at cohort entry, mean (SD) years | 55.4 (21.1) | 54.9 (21.1) | 64.2 (19.0) | 60.3 (20.4) |
| Year of cohort entry, N (%) | | | | |
| 2008 | 467 (12.4) | 4,290 (14.0) | 5,858 (12.4) | 10,615 (13.0) |
| 2009 | 862 (22.9) | 7,128 (23.3) | 11,173 (23.6) | 19,163 (23.4) |
| 2010 | 852 (22.6) | 6,851 (22.4) | 10,572 (22.3) | 18,275 (22.3) |
| 2011 | 885 (23.5) | 6,622 (21.6) | 10,172 (21.5) | 17,679 (21.6) |
| 2012 | 698 (18.5) | 5,746 (18.8) | 9,609 (20.3) | 16,053 (19.6) |
| Hispanic, N (%) | 300 (8.0) | 3,542 (11.6) | 836 (1.8) | 4,678 (5.7) |
| Race, N (%) | | | | |
| White | 2,807 (74.6) | 22,986 (75.0) | 32,676 (69.0) | 58,469 (71.5) |
| African American | 139 (3.7) | 2,605 (8.5) | 3,369 (7.1) | 6,113 (7.5) |
| Other | 83 (2.2) | 2,393 (7.8) | 546 (1.2) | 3,022 (3.7) |
| Unknown | 735 (19.5) | 2,653 (8.7) | 10,793 (22.8) | 14,181 (17.3) |
| GLU CPT code during baseline, N (%) yes [b] | 1,758 (46.7) | 11,583 (37.8) | 20,497 (43.3) | 33,838 (41.4) |
| GLU CPT code during follow-up, N (%) yes [b] | 1,711 (45.5) | 11,007 (35.9) | 21,756 (45.9) | 34,474 (42.2) |
| Number of unique medication classes dispensed during baseline, mean (SD) | 9.0 (4.7) | 9.2 (4.7) | 10.6 (5.0) | 10.0 (4.9) |
| Number of ambulatory healthcare visits during baseline, mean (SD) | 6.6 (6.5) | 11.1 (18.2) | 16.8 (17.7) | 14.2 (17.8) |
| Any emergency department visit during baseline, N (%) yes | 1,329 (35.3) | 14,053 (45.9) | 15,659 (33.0) | 31,041 (38.0) |
| Any hospitalization during baseline, N (%) yes | 944 (25.1) | 8,153 (26.6) | 12,329 (26.0) | 21,426 (26.2) |
| Any institutional stay during baseline, N (%) yes | 357 (9.5) | 2,617 (8.5) | 15,597 (32.9) | 18,571 (22.7) |
| Comorbidity score, mean (SD) [c, d] | 1.9 (2.0) | 1.6 (1.9) | 2.1 (2.5) | 1.9 (2.3) |
| Individual comorbidities, N (%) yes [c, d] | | | | |
| Alcohol abuse | 435 (11.6) | 3,391 (11.1) | 2,687 (5.7) | 6,513 (8.0) |
| Anemia | 312 (8.3) | 3,023 (9.9) | 8,741 (18.4) | 12,076 (14.8) |
| Cardiac arrhythmia | 384 (10.2) | 3,004 (9.8) | 7,703 (16.3) | 11,091 (13.6) |
| Coagulopathy | 95 (2.5) | 627 (2.0) | 1,479 (3.1) | 2,201 (2.7) |
| Heart failure, chronic | 251 (6.7) | 1,802 (5.9) | 6,230 (13.1) | 8,283 (10.1) |
| Dementia | 534 (14.2) | 2,872 (9.4) | 11,463 (24.2) | 14,869 (18.2) |
| Fluid/electrolyte disorder | 566 (15.0) | 3,224 (10.5) | 8,423 (17.8) | 12,213 (14.9) |
| HIV | 6 (0.2) | 134 (0.4) | 184 (0.4) | 324 (0.4) |
| Hypertension | 1,146 (30.4) | 10,285 (33.6) | 25,378 (53.6) | 36,809 (45.0) |
| Hemiplegia | 37 (1.0) | 310 (1.0) | 969 (2.0) | 1,316 (1.6) |
| Liver disease | 96 (2.6) | 846 (2.8) | 1,162 (2.5) | 2,104 (2.6) |
| Metastatic cancer | 28 (0.7) | 493 (1.6) | 685 (1.4) | 1,206 (1.5) |
| Psychosis | 2,769 (73.6) | 19,857 (64.8) | 23,165 (48.9) | 45,791 (56.0) |
| Pulmonary disease | 576 (15.3) | 4,635 (15.1) | 9,919 (20.9) | 15,130 (18.5) |
| Pulmonary circulation disorder | 97 (2.6) | 307 (1.0) | 1,018 (2.1) | 1,422 (1.7) |

| Variable | Data Partner Site | | | |
|---|---|---|---|---|
| | Site 1 (N = 3,764) | Site 2 (N = 30,637) | Site 3 (N = 47,384) | All Sites (N = 81,785) |
| PVD | 167 (4.4) | 1,683 (5.5) | 5,700 (12.0) | 7,550 (9.2) |
| Renal | 338 (9.0) | 2,271 (7.4) | 4,758 (10.0) | 7,367 (9.0) |
| Tumor | 178 (4.7) | 1,634 (5.3) | 3,706 (7.8) | 5,518 (6.7) |
| Weight loss | 112 (3.0) | 703 (2.3) | 1,648 (3.5) | 2,463 (3.0) |
| Individual comorbidities specific to test case [c] | | | | |
| Myocardial infarction, acute | 130 (3.5) | 1,208 (3.9) | 2,277 (4.8) | 3,615 (4.4) |
| Ischemic Stroke | 135 (3.6) | 875 (2.9) | 4,754 (10.0) | 5,764 (7.0) |
| Intracranial Hemorrhage | 47 (1.2) | 290 (0.9) | 670 (1.4) | 1,007 (1.2) |
| Osteoarthritis | 483 (12.8) | 3,913 (12.8) | 9,762 (20.6) | 14,158 (17.3) |
| Depression | 2,264 (60.1) | 17,482 (57.1) | 24,111 (50.9) | 43,857 (53.6) |

[a] Within 365 days after the cohort entry date
[b] CPT codes 80047, 80048, 80050, 80053, 80069, 82947, 83036, 83037
[c] Determined over the 183 days prior to the cohort entry date
[d] Gagne et al [69]

Less than half (40.5%) of the cohort had any baseline GLU result available; 46.4% had any follow-up GLU result available. Availability of follow-up GLU results differed by site (Table 37), from 31.9% at site 3 to 66.2% at site 2 to 68.2% at site 1. Random glucose accounted for 52.4% of the 68.2% with GLU available at site 1 and 43.2% of the 66.2% with GLU available at site 2, at site 3, 31.0% of the 31.9% with GLU available were random glucoses. Unique combinations of type(s) of GLU results are shown in Table 38 for baseline and follow-up periods. Patterns of available GLU result types appear similar in the two periods, with random glucose most common and HbA1c least common.

**Table 38. Availability of HbA1c, Fasting Glucose, and/or Random Glucose Laboratory Test Result Values during Baseline Period and within 365 Days after Initiation of Second Generation Antipsychotic**

| Laboratory Test Result Type | Laboratory Result Value Available during Baseline Period | Laboratory Result Value Available during 365 Day Follow-up Period |
|---|---|---|
| | Number of Individuals (%) | Number of Individuals (%) |
| None | 48,668 (59.5) | 43,827 (53.6) |
| HbA1c only | 376 (0.5) | 541 (0.7) |
| Fasting glucose only | 3,950 (4.8) | 6,141 (7.5) |
| Random glucose only | 22,852 (27.9) | 21,091 (25.8) |
| HbA1c and Fasting glucose | 681 (0.8) | 1,386 (1.7) |
| HbA1c and random glucose | 2,378 (2.9) | 3,769 (4.6) |
| Fasting and random glucose | 2,324 (2.8) | 3,691 (4.5) |
| HbA1c and fasting and random glucose | 556 (0.7) | 1,339 (1.6) |
| Any HbA1c, fasting glucose, or random glucose | 33,117 (40.5) | 37,958 (46.4) |

The number and proportion of individuals with a diabetes outcome indicator identified from a coded diagnosis or medication dispensing versus a diabetes diagnosis, medication dispensing or GLU result are shown in Table 39.

**Table 39. Outcomes Identification Test Case 1, Numbers of Diabetes Outcomes Detected among Individuals Initiating a Second Generation Antipsychotic Agent using Diagnosis, Medication, and Laboratory Results Criteria**

| Second Generation Antipsychotic Agent | Diabetes Outcome Indicator | | | | |
|---|---|---|---|---|---|
| | Coded Diagnosis or Antidiabetic Medication | Coded Diagnosis, Antidiabetic Medication or HbA1c ≥ 6.5 | Coded Diagnosis, Antidiabetic Medication, or Fasting Glucose ≥ 126 | Coded Diagnosis, Antidiabetic Medication, or Random Glucose ≥ 200 | Coded Diagnosis, Antidiabetic Medication, HbA1c ≥ 6.5, Fasting Glucose ≥ 126, or Random Glucose ≥ 200 |
| Aripiprazole (N = 14,333) | 645 (4.5) | 666 (4.7) | 698 (4.9) | 673 (4.7) | 744 (5.2) |
| Olanzapine (N = 8,864) | 453 (5.1) | 468 (5.3) | 488 (5.5) | 530 (6.0) | 574 (6.5) |
| Quetiapine (N = 39,353) | 1,953 (5.0) | 2,029 (5.2) | 2,084 (5.3) | 2,235 (5.7) | 2,411 (6.1) |
| Risperidone (N = 19,235) | 1,107 (5.8) | 1,144 (6.0) | 1,161 (6.0) | 1,181 (6.1) | 1,259 (6.5) |
| Total (N = 81,785) | 4,158 (5.1) | 4,307 (5.3) | 4,431 (5.4) | 4,619 (5.6) | 4,988 (6.1) |

Moving from using only a coded diagnosis or medication dispensing to also using any GLU result values, the number of preliminary diabetes outcomes identified increased from 4,158 to 4,988, an absolute 1.0% increase (17% relative increase). For each SGA, the largest increase was due to considering fasting or random glucose, not HbA1c. Inclusion of any GLU result yielded a similar absolute increase in diabetes outcomes identified for aripiprazole (0.7%) and risperidone (0.7%) and a greater absolute increase in diabetes outcomes identified for olanzapine (1.4%) and quetiapine (1.1%).

Table 40 provides results for the impact of GLU results on the time to preliminary indicator of diabetes when GLU results were and were not considered. Median was shorter than mean time to diabetes indicator across all SGAs and all indicators, and did not differ substantially across diabetes indicators. Time to diabetes indicator was longer for aripiprazole (median 139.5 days when diagnosis, medication dispensing, and any GLU result value was considered) than for olanzapine (median 116 days), quetiapine (median 126 days), or risperidone (median 118 days). Overall, including GLU result values did not identify diabetes outcomes earlier, but did identify additional outcomes.

**Table 40. Time to Preliminary Diabetes Indicator when GLU Results were and were not considered**

| Diabetes Indicators and Second Generation Antipsychotic | Number of Individuals | Number of Days | |
|---|---|---|---|
| | | Mean (SD) | Median [a] |
| Diagnosis by ICD9 or Meds | | | |
| Aripiprazole | 645 | 156.2 (104.1) | 144.0 |
| Olanzapine | 453 | 135.0 (102.7) | 114.0 |
| Quetiapine | 1953 | 143.7 (106.8) | 126.0 |
| Risperidone | 1,107 | 138.3 (105.0) | 116.0 |
| Diagnosis by ICD9, Meds or HbA1c ≥ 6.5 | | | |
| Aripiprazole | 666 | 154.5 (104.3) | 139.0 |
| Olanzapine | 468 | 134.1 (102.5) | 114.0 |
| Quetiapine | 2,029 | 143.7 (106.5) | 126.0 |
| Risperidone | 1,144 | 138.9 (105.0) | 118.0 |
| Diagnosis by ICD9, Meds or fasting glucose ≥ 126 | | | |
| Aripiprazole | 698 | 155.9 (105.7) | 142.0 |
| Olanzapine | 488 | 137.3 (103.8) | 117.5 |
| Quetiapine | 2,084 | 144.1 (107.1) | 126.0 |
| Risperidone | 1,161 | 138.5 (105.4) | 116.0 |
| Diagnosis by ICD9, Meds or random glucose ≥ 200 | | | |
| Aripiprazole | 673 | 155.2 (103.9) | 144.0 |
| Olanzapine | 530 | 135.5 (103.8) | 114.0 |
| Quetiapine | 2,235 | 145.1 (107.2) | 127.0 |
| Risperidone | 1,181 | 140.0 (105.0) | 119.0 |
| Diagnosis by ICD9, Meds, HbA1c ≥ 6.5, fasting glucose ≥ 126 or random glucose ≥ 200 | | | |
| Aripiprazole | 744 | 154.0 (105.6) | 139.5 |
| Olanzapine | 574 | 136.4 (104.2) | 116.0 |
| Quetiapine | 2,411 | 145.0 (107.2) | 126.0 |
| Risperidone | 1,259 | 139.7 (105.0) | 118.0 |
| [a] Minimum and maximum days for all cells were 0 and 365, respectively | | | |

### c. Outcome Detection Test Case 1, predictors of missing blood glucose or glycosylated hemoglobin laboratory test results values

***Diabetes outcome rates with and without considering GLU laboratory result values***

Crude preliminary diabetes outcome rates per person year are shown in Table 41. The rate for each site was first calculated only using diagnosis codes or medication dispensings and second using diagnosis codes, medication dispensings, or GLU result values.

**Table 41. Crude Diabetes Outcome Rates per Person Year by Site Based on Diagnosis and Medication Use Indicator(s) and by Diagnosis, Medication, and GLU (Glucose or HbA1c) Result Values**

| Site | Crude Diabetes Outcome Rate Per Person-Year | |
|---|---|---|
| | Determined Using ICD-9 Codes or Antidiabetic Medication Dispensing | Determined Using ICD-9 Codes, Antidiabetic Medication Dispensing, and GLU Laboratory Result Values |
| Site 1 (N = 3,764) | 0.024 (0.019, 0.029) | 0.038 (0.032, 0.045) |
| Site 2 (N = 30,637) | 0.018 (0.016, 0.019) | 0.039 (0.037, 0.041) |
| Site 3 (N = 47,384) | 0.078 (0.076, 0.081) | 0.082 (0.079, 0.084) |

At all sites, the diabetes outcome rate increased when diagnosis codes, medication dispensings, and GLU results were considered versus only considering diagnosis codes and medication dispensings. The largest increase was at site 2 (from 0.018 to 0.039). The rate increased the least at site 3 (from 0.078 to 0.082). Reasons the rate increases differ across sites potentially include coding practice variations (i.e., site 3 has a crude outcome rate substantially higher than either of the other sites based on coded diagnosis or medication dispensing) and that at site 3 only 31.9% of individuals had a follow-up GLU, whereas sites 1 and 2 had follow-up GLU for at least 66.2% of the individuals they contributed.

To assist in understanding the contribution of missing GLU result values, at site 3 we computed outcome rates among individuals who did (N = 12,990) and did not (N = 34,394) have a *baseline GLU procedure claim* (CPT code) regardless of whether or not they had any follow-up GLU result available. The rate of diabetes outcome was lower in the subgroup that had a baseline GLU procedure claim than in the subgroup without a baseline GLU procedure claim (Table 42). One possible explanation for this finding could be that this is an "artifact of the exclusion criteria. " The subgroup without baseline GLU procedure claims likely included some patients with undiagnosed pre-existing elevated GLU (or diabetes) that appeared (erroneously) to be incident diabetes when GLU laboratory testing was completed during follow-up. That is, patients who had baseline GLU laboratory testing result values compatible with diabetes were excluded from the cohort, while those without any baseline GLU laboratory testing (or baseline GLU result values) did not have the opportunity to be excluded from the cohort on the basis of GLU laboratory result values (i.e., differential exclusion at baseline).

**Table 42. Crude Diabetes Outcome Rates per Person Year at Site 3 Based on Diagnosis and Medication Use Indicator(s) or by Diagnosis, Medication, and GLU Follow-up Result Values Stratified by Presence of a GLU Laboratory Procedure Claim in the Baseline Period**

| Site 3, Stratified by Presence or Absence of a Baseline Laboratory Procedure Claim | Crude Diabetes Outcome Rate per Person-Year | |
|---|---|---|
| | Determined Using ICD-9 Codes or Antidiabetic Medication Dispensing | Determined Using ICD-9 Codes, Antidiabetic Medication Dispensing, and GLU Laboratory Result Values |
| Overall (N = 47,384) | 0.078 (0.076, 0.081) | 0.082 (0.079, 0.084) |
| Baseline GLU laboratory procedure claim present (N = 12,990) | 0.073 (0.07, 0.076) | 0.075 (0.072, 0.078) |
| Baseline GLU laboratory procedure claim not present (N = 34,394) | 0.092 (0.086, 0.097) | 0.098 (0.093, 0.104) |

Crude diabetes outcome rates are presented in Table 43 for each site, stratified by whether or not any follow-up GLU result value was available. At all sites, the outcome rate is much higher among individuals with follow-up GLU results available (regardless of whether the outcome is determined using GLU indicators), suggesting bias in selecting which individuals receive GLU evaluation (e.g., purposeful evaluation). This finding strongly supports the possibility that clinicians selectively choose patients for glucose-related monitoring who they believe are at higher risk of diabetes.

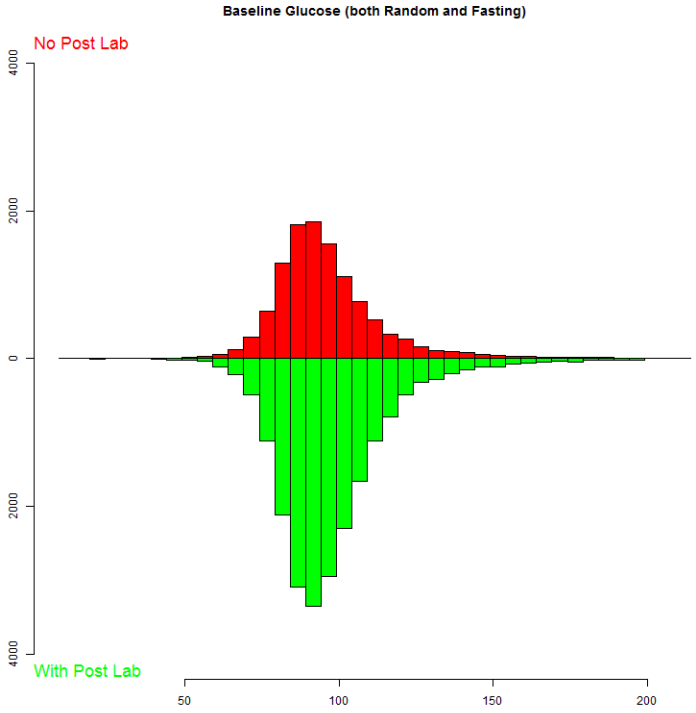
**Table 43. Crude Diabetes Outcome Rates per Person Year by Site Based on Diagnosis and Medication Use Indicator(s) and by Diagnosis, Medication, and GLU Result Value(s) in the Follow-Up Period**

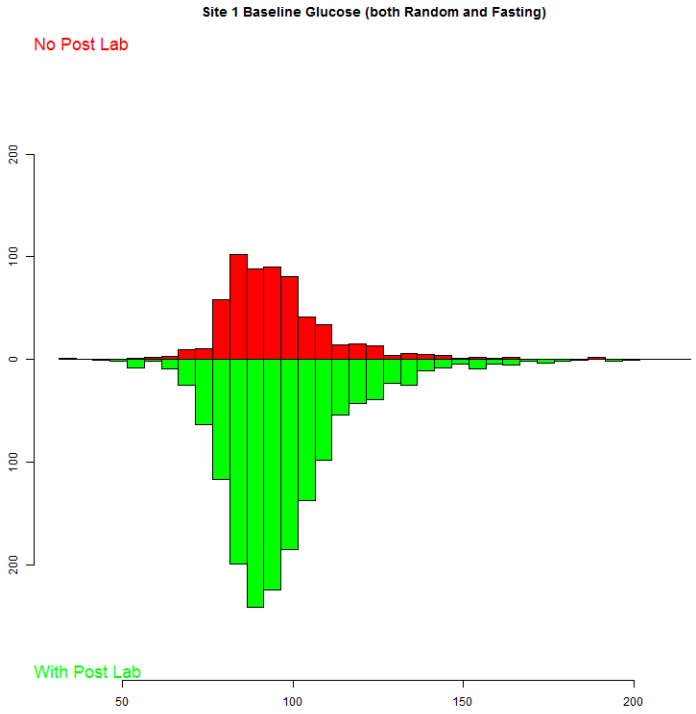
| Site and Availability of GLU Laboratory Result Value During Follow-up Period | Crude Diabetes Outcome Rate per Person-Year | |
|---|---|---|
| | Determined Using ICD-9 Codes or Antidiabetic Medication Dispensing | Determined Using ICD-9 Codes, Antidiabetic Medication Dispensing, and GLU Laboratory Results Values |
| Site 1 (N = 3,764) | | |
| Available in follow-up period (N = 2,567) | 0.033 (0.027, 0.041) | 0.055 (0.046, 0.065) |
| Not available in follow-up period (N = 1,197) | 0.004 (0.002, 0.01) | 0.004 (0.002, 0.01) |
| Site 2 (N = 30,637) | | |
| Available in follow-up period (N = 20,268) | 0.025 (0.023, 0.028) | 0.058 (0.055, 0.061) |
| Not available in follow-up period (N = 10,369) | 0.003 (0.002, 0.004) | 0.003 (0.002, 0.004) |
| Site 3 (N = 47,384) | | |
| Available in follow-up period (N = 15,123) | 0.119 (0.113, 0.124) | 0.13 (0.124, 0.136) |
| Not available in follow-up period (N = 32,261) | 0.06 (0.057, 0.062) | 0.06 0.057, 0.062) |

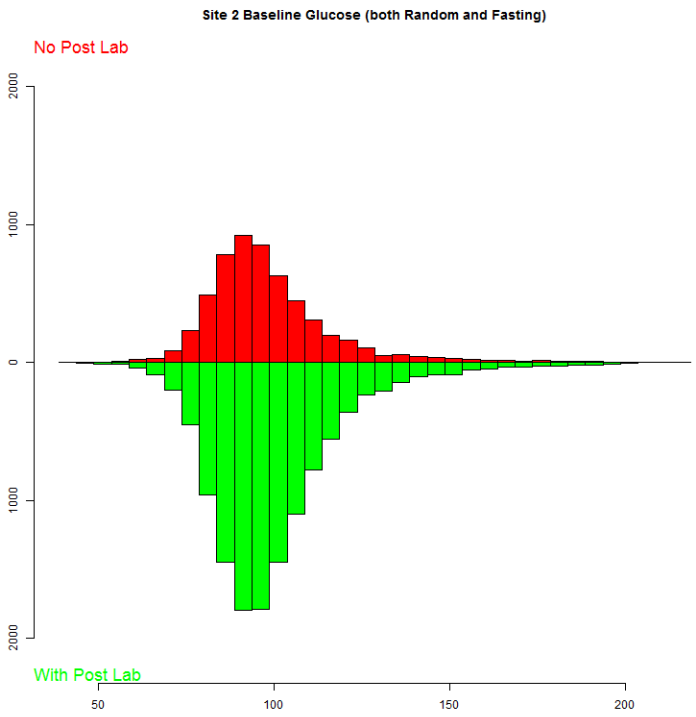
***Baseline GLU laboratory result values as predictors of missing follow-up GLU laboratory result values***

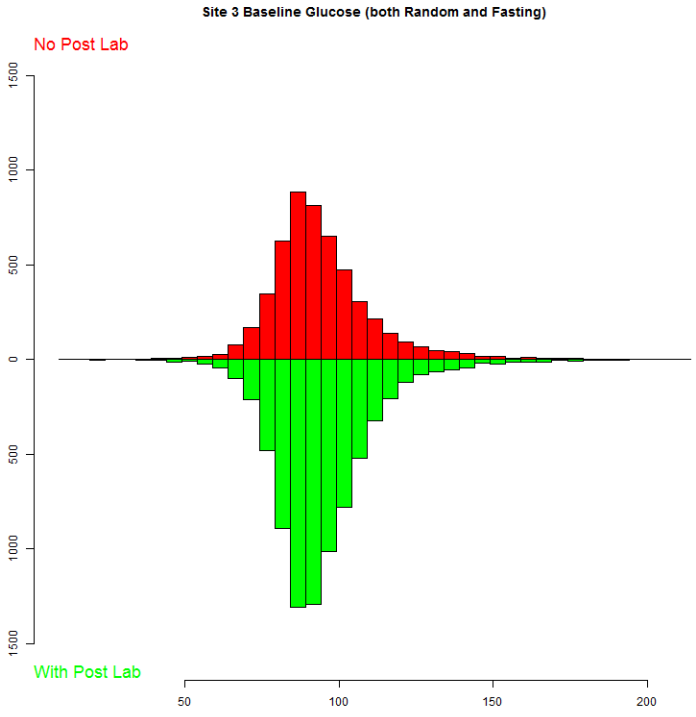
The Workgroup explored whether the *value of a baseline GLU result* predicted missingness of a *follow-up GLU result*. For individuals who had a baseline GLU result, we examined the distribution of baseline actual GLU result values, stratified by whether or not they had corresponding follow-up GLU result values. Below are mirrored histograms for baseline random or fasting glucose result values (**Figure 8**, Panel A) and baseline HbA1c result values (**Figure 8**, Panel B), overall and by site. In each figure the distribution of the baseline result value is displayed between individuals without a follow-up GLU result value ("no post lab" in red font on the top half of each figure) and individuals with a follow-up GLU result value ("with post lab" in green font on the bottom half of each figure).

**Figure 8. Panel A: Distribution of Baseline Random or Fasting Glucose Result Values Stratified by Whether a Random or Fasting Glucose Result Value was also available in the Follow-up Period Overall and by Site**

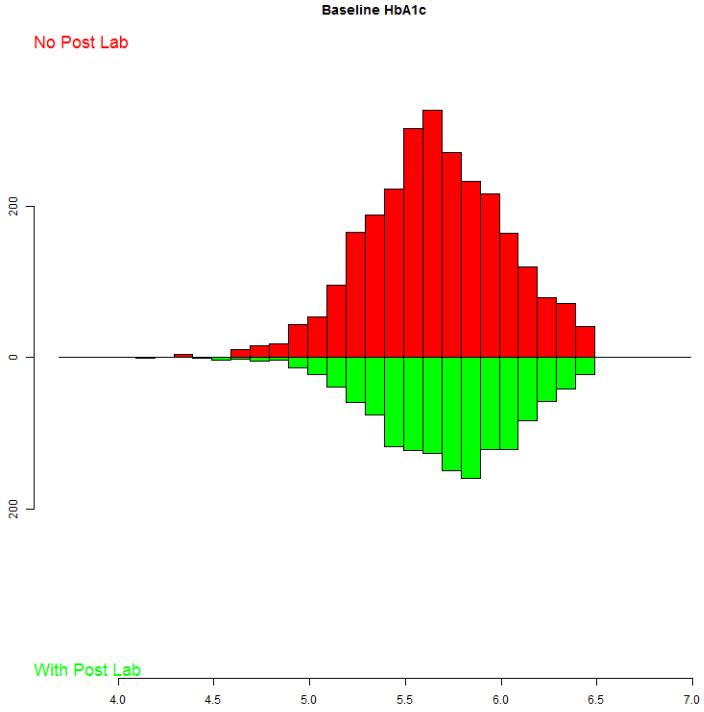


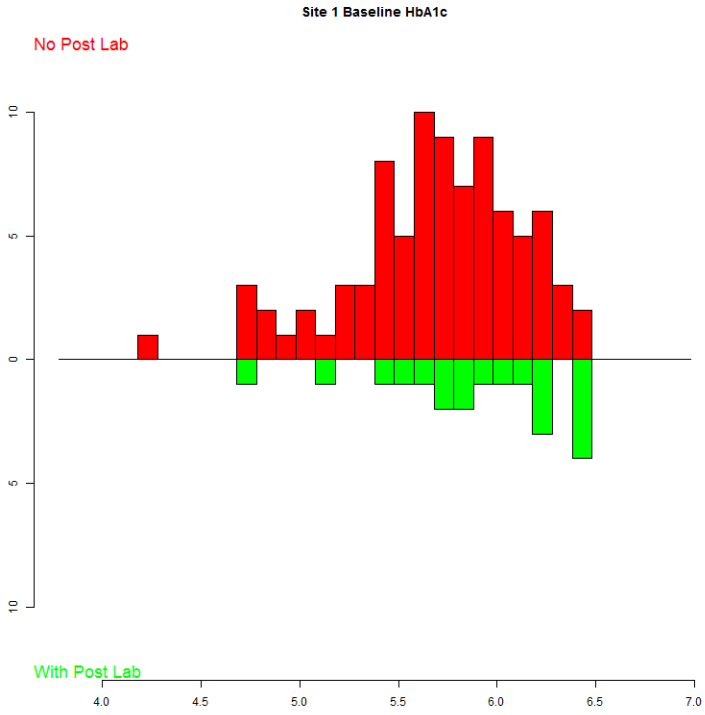



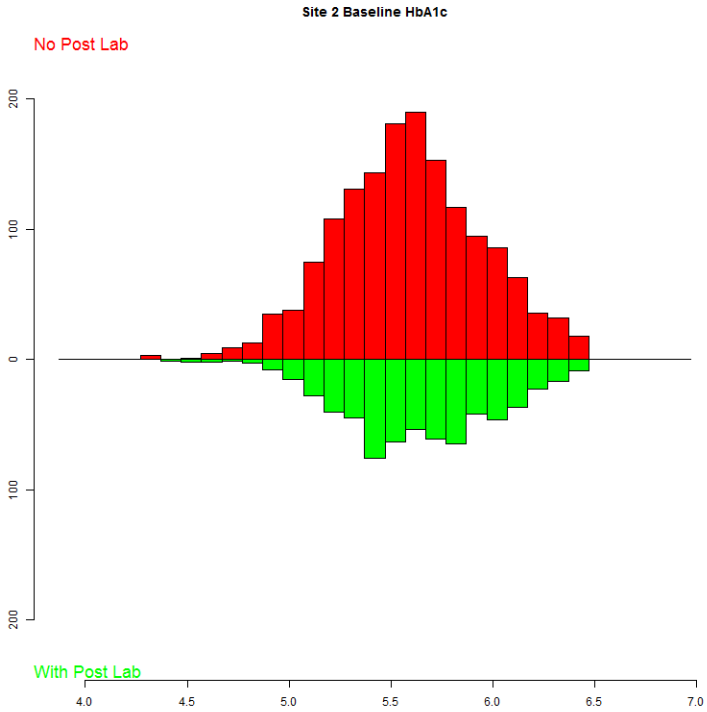

Site 2 Baseline Glucose (both Random and Fasting)
No Post Lab
2000
1000
0
1000
2000
With Post Lab
50
100
150
200

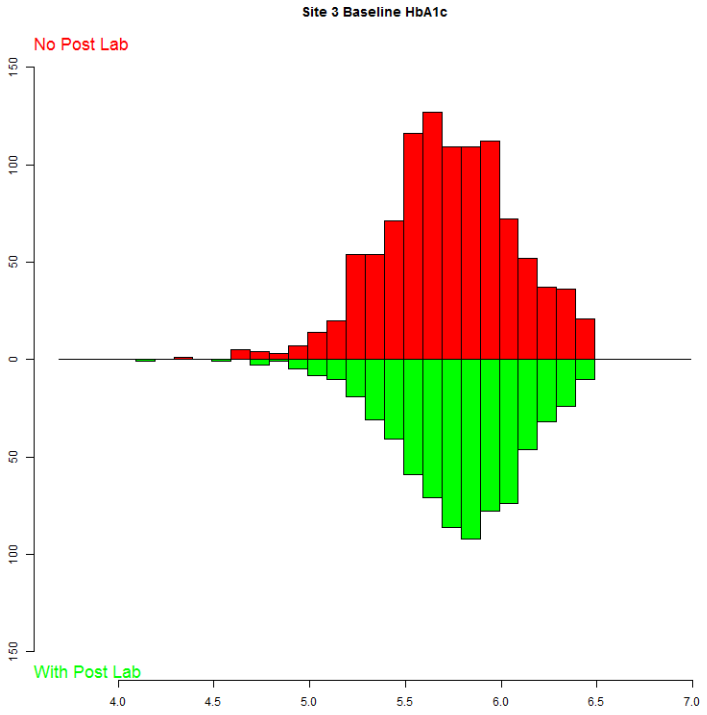

Site 3 Baseline Glucose (both Random and Fasting)
No Post Lab
1500
1000
500
0
500
1000
1500
With Post Lab
50
100
150
200

**Panel B: Distribution of Baseline Glycosylated Hemoglobin Result Values Stratified by Whether a Glycosylated Hemoglobin Result Value was also available in the Follow-up Period Overall and by Site**




Site 2 Baseline HbA1c
No Post Lab
200
100
0
100
200
With Post Lab
4.0
4.5
5.0
5.5
6.0
6.5
7.0


Site 3 Baseline HbA1c
No Post Lab
150
100
50
0
50
100
150
With Post Lab
4.0
4.5
5.0
5.5
6.0
6.5
7.0

For individuals with a baseline random or fasting glucose result value (Figure 8, Panel A), regardless of whether sites are considered together or individually, there is no obvious relationship between the value of the baseline glucose result and missingness of a follow-up glucose result (the shapes of the no post lab and with post lab histograms are similar). However, for individuals with a baseline HbA1c result value (Figure 8, Panel B), it appears a higher baseline HbA1c result value predicts having a follow-up HbA1c result value available (Figure 8, first graphic in Panel B, where the mode in the post lab subgroup is 5.9% and the mode in the no post lab subgroup is 5.6%). Site 1 has few HbA1c observations; that histogram should be interpreted cautiously. Site 2 histograms for post lab and no post lab subgroups are similar. Site 3 has a greater shift to the right in the post lab subgroup compared to the no post lab subgroup.

***Relationships between baseline characteristics and missing follow-up GLU result values***

Logistic regression was used to investigate relationships between baseline variables and the probability of missing follow-up GLU results. For these models, the outcome was no observed follow-up GLU result values (**Table 44**).

In the fully-adjusted analysis of sites combined, not having a baseline GLU result available was strongly associated with not having a follow-up GLU result available (OR 0.29). This association was only at site 3 (OR 0.19 [0.18, 0.20]; site 1: OR 0.93 [0.79, 1.09]; site 2: OR 1.07 [1.01, 1.13]), the site that had the lowest percentage of the cohort with GLU results available. A combination of system and patient characteristics are potential contributors to this finding: Patients are likely to consistently use the same commercial laboratory to have their blood drawn for GLU assessment; if they consistently use a laboratory vendor contracted to provide laboratory result values to the health insurer (site 3), then those GLU result values would be available for both the baseline and follow-up period (the converse would also be true).

Olanzapine (OR 0.93), quetiapine (OR 0.87), or risperidone (OR 0.91) all had lower likelihood of missing follow-up GLU results than did the reference SGA (aripiprazole). Other associations are also shown in **Table 44**.

**Table 44. Outcome Detection Test Case 1, Initiation of Second Generation Antipsychotics and Diabetes Outcome: Adjusted Logistic Regression Models assessing Associations with Missing Follow-up GLU Results, Overall and by Site**

| Characteristic | Associations with Missing Follow-up GLU: Adjusted Odds Ratios (95% CI) [a] | | | |
|---|---|---|---|---|
| | Data Partner Site | | | |
| | All Sites Combined | Site 1 | Site 2 | Site 3 |
| SGA, Aripiprazole reference | | | | |
| Olanzapine | 0.93 (0.87, 0.98) | 0.72 (0.50, 1.04) | 0.78 (0.71, 0.87) | 1.02 (0.94, 1.11) |
| Quetiapine | 0.87 (0.84, 0.91) | 0.84 (0.66, 1.07) | 0.95 (0.88, 1.02) | 0.96 (0.90, 1.02) |
| Risperidone | 0.91 (0.86, 0.95) | 0.69 (0.55, 0.88) | 0.78 (0.72, 0.85) | 0.97 (0.91, 1.04) |
| Sex, male vs. female | 1.08 (1.04, 1.11) | 1.08 (0.93, 1.26) | 0.99 (0.94, 1.04) | 1.09 (1.04, 1.14) |
| Age (per 10 years) | 0.99 (0.98, 1.00) | 0.81 (0.77, 0.85) | 0.84 (0.83, 0.86) | 1.05 (1.03, 1.07) |
| Any Baseline GLU | 0.29 (0.28, 0.29) | 0.93 (0.79, 1.09) | 1.07 (1.01, 1.13) | 0.19 (0.18, 0.20) |
| Year of cohort entry, 2008 reference | | | | |
| 2009 | 0.94 (0.89, 0.99) | 0.96 (0.75, 1.23) | 0.95 (0.88, 1.03) | 0.85 (0.79, 0.92) |
| 2010 | 0.91 (0.86, 0.96) | 0.89 (0.69, 1.15) | 0.96 (0.88, 1.04) | 0.82 (0.76, 0.88) |
| 2011 | 0.83 (0.79, 0.88) | 0.77 (0.60, 1.00) | 0.95 (0.87, 1.03) | 0.68 (0.63, 0.74) |
| 2012 | 0.87 (0.83, 0.92) | 1.01 (0.78, 1.31) | 0.99 (0.91, 1.08) | 0.69 (0.64, 0.75) |
| Hispanic, no/unknown reference | 0.44 (0.41, 0.47) | 0.91 (0.68, 1.22) | 0.71 (0.65, 0.78) | 0.36 (0.30, 0.42) |
| Race, unknown reference | | | | |
| African American | 0.52 (0.49, 0.56) | 0.99 (0.66, 1.49) | 0.62 (0.54, 0.70) | 0.65 (0.59, 0.72) |
| White | 0.57 (0.54, 0.59) | 0.83 (0.68, 1.01) | 0.67 (0.61, 0.74) | 0.74 (0.69, 0.79) |
| Other | 0.41 (0.37, 0.45) | 0.77 (0.47, 1.27) | 0.59 (0.52, 0.67) | 0.75 (0.61, 0.92) |
| Number of unique medication classes dispensed during baseline | 0.97 (0.97, 0.97) | 0.93 (0.91, 0.95 ) | 0.95 (0.94 , 0.95 ) | 0.97 (0.96 , 0.97 ) |
| Number of ambulatory healthcare visits during baseline | 1.00 (1.00, 1.01) | 0.99 (0.97, 1.00 ) | 1.00 (1.00 , 1.00 ) | 1.00 (1.00 , 1.00 ) |
| Emergency department visit during baseline (Y vs. N) | 1.05 (1.01, 1.08) | 1.07 (0.90, 1.27) | 0.86 (0.80, 0.91) | 1.20 (1.14, 1.26) |
| Hospitalization during baseline (Y vs. N) | 1.01 (0.97, 1.06) | 0.93 (0.74, 1.16) | 0.93 (0.86, 1.00) | 1.16 (1.08, 1.24) |
| Institutional stay during baseline (Y vs. N) | 1.86 (1.78, 1.94) | 0.89 (0.65, 1.22) | 2.20 (1.98, 2.44) | 1.29 (1.22, 1.36) |
| Alcohol abuse (Y vs. N) | 0.86 (0.82, 0.92) | 1.00 (0.79, 1.27) | 0.98 (0.90, 1.06) | 0.98 (0.89, 1.07) |
| Anemia (Y vs. N) | 1.12 (1.07, 1.17) | 1.06 (0.77, 1.46) | 0.91 (0.83, 1.01) | 1.10 (1.04, 1.17) |
| Cardiac arrhythmia (Y vs. N) | 0.97 (0.92, 1.02) | 0.83 (0.61, 1.12) | 0.93 (0.84, 1.03) | 1.00 (0.94, 1.07) |
| Coagulopathy (Y vs. N) | 0.92 (0.83, 1.01) | 0.79 (0.44, 1.39) | 1.10 (0.91, 1.34) | 0.88 (0.78, 1.00) |
| Heart failure, chronic (Y vs. N) | 1.13 (1.06, 1.20) | 1.13 (0.77, 1.67) | 1.21 (1.05, 1.38) | 1.07 (1.00, 1.16) |
| Dementia (Y vs. N) | 1.45 (1.38, 1.51) | 1.26 (0.97, 1.64) | 1.08 (0.98, 1.19) | 1.38 (1.30, 1.46) |
| Fluid/electrolyte disorder (Y vs. N) | 1.05 (1.00, 1.11) | 0.80 (0.61, 1.04) | 0.93 (0.84, 1.03) | 1.08 (1.01, 1.16) |
| HIV (Y vs. N) | 0.53 (0.41, 0.68) | 0.00 (0.00, ) | 0.44 (0.27, 0.70) | 0.52 (0.37, 0.73) |
| Hypertension (Y vs. N) | 1.01 (0.97, 1.05) | 0.96 (0.78, 1.18) | 0.85 (0.79, 0.91) | 0.90 (0.86, 0.95) |
| Hemiplegia (Y vs. N) | 1.10 (0.97, 1.25) | 0.96 (0.43, 2.14) | 0.85 (0.65, 1.13) | 1.30 (1.10, 1.54) |
| Liver disease (Y vs. N) | 0.78 (0.71, 0.86) | 0.74 (0.43, 1.27) | 0.81 (0.69, 0.95) | 0.74 (0.65, 0.85) |
| Metastatic cancer (Y vs. N) | 1.71 (1.49, 1.97) | 2.07 (0.79, 5.38) | 2.76 (2.23, 3.43) | 1.59 (1.30, 1.95) |
| Psychosis (Y vs. N) | 0.75 (0.72, 0.77) | 0.74 (0.61, 0.90) | 0.74 (0.70, 0.79) | 0.86 (0.82, 0.90) |
| Pulmonary disease (Y vs. N) | 1.03 (0.99, 1.07) | 1.12 (0.90, 1.40) | 0.98 (0.91, 1.06) | 0.97 (0.91, 1.02) |
| Pulmonary circulation disorder (Y vs. N) | 1.11 (0.99, 1.26) | 1.01 (0.57, 1.78) | 1.29 (0.98, 1.71) | 1.20 (1.03, 1.40) |
| PVD (Y vs. N) | 1.07 (1.01, 1.13) | 1.53 (1.02, 2.28) | 1.05 (0.92, 1.20) | 0.99 (0.92, 1.06) |
| Renal (Y vs. N) | 0.76 (0.72, 0.81) | 1.33 (0.98, 1.81) | 0.97 (0.87, 1.09) | 0.80 (0.74, 0.86) |
| Tumor (Y vs. N) | 0.95 (0.89, 1.02) | 0.81 (0.53, 1.24) | 1.19 (1.05, 1.36) | 0.87 (0.79, 0.94) |
| Weight loss (Y vs. N) | 1.01 (0.92, 1.12) | 1.48 (0.90, 2.45) | 0.99 (0.82, 1.20) | 1.24 (1.08, 1.41) |
| Myocardial infarction, acute (Y vs. N) | 0.83 (0.76, 0.89) | 1.29 (0.82, 2.01) | 0.91 (0.78, 1.06) | 0.89 (0.80, 0.99) |
| Ischemic Stroke (Y vs. N) | 1.09 (1.02, 1.17) | 1.35 (0.85, 2.15) | 0.91 (0.76, 1.09) | 0.95 (0.88, 1.03) |
| Intracranial Hemorrhage (Y vs. N) | 0.85 (0.74, 0.99) | 0.86 (0.41, 1.81) | 0.67 (0.50, 0.90) | 1.04 (0.86, 1.27) |

| Characteristic | Associations with Missing Follow-up GLU: Adjusted Odds Ratios (95% CI) [a] | | | |
|---|---|---|---|---|
| | Data Partner Site | | | |
| | All Sites Combined | Site 1 | Site 2 | Site 3 |
| Osteoarthritis (Y vs. N) | 0.95 (0.91, 0.99) | 1.11 (0.86, 1.43) | 0.94 (0.86, 1.02) | 0.92 (0.87, 0.97) |
| Depression (Y vs. N) | 0.94 (0.91, 0.98) | 1.14 (0.96, 1.35) | 1.03 (0.97, 1.09) | 0.91 (0.86, 0.95) |
| [a] Adjusted for all characteristics shown | | | | |

### d. Outcomes Detection Test Case 1, associations between individual SGA or baseline GLU availability and outcome with differing diabetes outcome indicators

We assessed the effect of the various outcome definitions on relationships between individual SGAs and the first preliminary indicator of diabetes for all sites combined and stratified by site using Cox proportional hazards models. Aripiprazole was the referent SGA for each comparison (Table 45). We also report HR for the availability of a baseline GLU result (Table 45).

The HR estimates for the all sites combined models, regardless of how the outcome was defined, were greater for olanzapine, quetiapine, and risperidone than for aripiprazole, but all CI contained 1. We did not detect a difference in risk of diabetes within 365 days after SGA initiation by specific SGA. When comparing the two most extreme diabetes outcome indicator definitions (Models 1 and 5), the largest difference in HR was for olanzapine (HR 1.02 versus 1.07), but all CI overlap; it does not appear this difference is meaningful.

The HR estimates for the all sites combined models for having any baseline GLU result available, regardless of how the outcome was defined, were all greater than 1, but the CI contained 1 in Models 1 and 3. In Model 2 (diagnosis codes, medication dispensing, or elevated HbA1c), Model 4 (diagnosis codes, medication dispensing, or elevated random glucose), and Model 5 (diagnosis codes, medication dispensing, or elevated any GLU), having a baseline GLU result available was associated with an increased risk of diabetes after SGA initiation. It is beyond the scope of the current work to determine whether this finding is a function of selection bias in ordering and performing baseline glucose-related assessment preferentially in patients perceived to be at higher risk of diabetes after SGA initiation. The differences in HR between models 1 and 5 (the two most extreme outcome definitions) for an available baseline GLU result were not meaningful (HR 1.07 versus 1.12, with CI that overlapped).

**Table 45. Outcome Detection Test Case 1, Initiation of Second Generation Antipsychotics and Diabetes Outcome: Associations between Second Generation Antipsychotic or Baseline GLU Result Availability and Diabetes Outcome with Differing Diabetes Outcome Indicators**

| Second Generation Antipsychotic Agent | Associations with Diabetes Outcomes: Adjusted Hazard Ratio (95% CI) | | | | |
|---|---|---|---|---|---|
| | Model 1: Diagnosis Code or Antidiabetic Medication Dispensing | Model 2: Diagnosis Code, Antidiabetic Medication Dispensing, or HbA1c ≥ 6.5 | Model 3: Diagnosis Code, Antidiabetic Medication Dispensing, or Fasting Glucose ≥ 126 | Model 4: Diagnosis Code, Antidiabetic Medication Dispensing, or Random Glucose ≥ 200 | Model 5: Diagnosis Code, Antidiabetic Medication Dispensing, or Any GLU (HbA1c ≥ 6.5, Fasting Glucose ≥ 126, Random Glucose ≥ 200 |
| All Sites, aripiprazole reference | | | | | |
| Olanzapine | 1.02 (0.90, 1.15) | 1.02 (0.90, 1.15) | 1.02 (0.90, 1.15) | 1.08 (0.96, 1.22) | 1.07 (0.96, 1.20) |
| Quetiapine | 1.03 (0.94, 1.13) | 1.03 (0.94, 1.13) | 1.01 (0.93, 1.11) | 1.06 (0.96, 1.16) | 1.04 ( 0.95, 1.13) |
| Risperidone | 1.03 (0.93, 1.15) | 1.04 (0.94, 1.15) | 1.03 (0.93, 1.13) | 1.02 (0.92, 1.13) | 1.01 (0.92, 1.11) |
| | | | | | |
| Any baseline GLU result available | 1.07 (1.00, 1.15) | 1.09 (1.02, 1.16) | 1.07 (1.00, 1.14) | 1.13 (1.06, 1.20) | 1.12 (1.06, 1.20) |
| | | | | | |
| Site Specific, aripiprazole reference | | | | | |
| Site 1 | | | | | |
| Olanzapine | 0.54 (0.14, 2.04) | 0.56 (0.15, 2.10) | 0.89 (0.30, 2.63) | 0.91 (0.31, 2.71) | 1.20 (0.46, 3.13) |
| Quetiapine | 0.92 (0.44, 1.93) | 1.18 (0.58, 2.43) | 1.11 (0.55, 2.21) | 1.34 (0.68, 2.65) | 1.53 (0.80, 2.90) |
| Risperidone | 1.18 (0.57, 2.44) | 1.25 (0.61, 2.55) | 1.26 (0.64, 2.51) | 1.34 (0.68, 2.66) | 1.43 (0.75, 2.73) |
| | | | | | |
| Any baseline GLU result available | 1.13 (0.69, 1.84) | 1.20 (0.76, 1.90) | 1.22 (0.78, 1.91) | 1.10 (0.72, 1.68) | 1.18 (0.80, 1.75) |
| | | | | | |
| Site 2 | | | | | |
| Olanzapine | 0.79 (0.55, 1.12) | 0.84 (0.61, 1.17) | 0.86 (0.65, 1.14) | 1.08 (0.82, 1.42) | 1.05 (0.83, 1.32) |
| Quetiapine | 0.90 (0.70, 1.15) | 0.90 (0.71, 1.14) | 0.86 (0.70, 1.05) | 1.00 (0.81, 1.25) | 0.92 (0.77, 1.10) |
| Risperidone | 0.88 (0.66, 1.19) | 0.95 (0.72, 1.25) | 0.88 (0.69, 1.12) | 0.93 (0.72, 1.21) | 0.90 (0.73, 1.11) |
| | | | | | |
| Any baseline GLU result available | 0.99 (0.80, 1.22) | 0.99 (0.82, 1.21) | 1.02 (0.85, 1.21) | 0.96 (0.80, 1.14) | 0.97 (0.84, 1.13) |
| | | | | | |
| Site 3 | | | | | |
| Olanzapine | 1.06 (0.93, 1.21) | 1.05 (0.92, 1.20) | 1.06 (0.93, 1.21) | 1.07 (0.94, 1.22) | 1.06 (0.93, 1.21) |
| Quetiapine | 1.05 (0.95, 1.16) | 1.05 (0.95, 1.16) | 1.05 (0.95, 1.17) | 1.05 (0.95, 1.17) | 1.06 (0.95, 1.17) |
| Risperidone | 1.05 (0.93, 1.17) | 1.04 (0.93, 1.17) | 1.05 (0.94, 1.18) | 1.04 (0.93, 1.16) | 1.04 (0.93, 1.16) |
| | | | | | |
| Any baseline GLU result available | 1.08 (1.01, 1.17) | 1.10 (1.02, 1.18) | 1.09 (1.01, 1.17) | 1.11 (1.04, 1.20) | 1.13 (1.05, 1.21) |

In the site-specific outcome models, comparing HR from different model definitions, the HR vary most widely at site 1, but CIs are wide and all contain 1. At sites 2 and 3, again all HR include 1. Within each outcome definition modeled, relatively large site differences are observed. For example, using the most liberal outcome definition (Model 5), the HR for quetiapine is 0.90 (0.73, 1.11) at site 2 and 1.06 (0.95, 1.17) at site 3. While these CIs overlap, the HR differ, and the lower limit of the CI at site 3 is greater than the HR estimate at site 2.

#### e. Summary of Outcomes Detection Test Case 1

Overall, less than half of the cohort had any baseline or follow-up GLU result value available. Availability of follow-up GLU results differed by site, with the large national insurer site having 31.9% with follow-up GLU and the smaller and larger integrated healthcare delivery system sites having 66.2% to 68.2% with follow-up GLU. Including GLU results did not identify diabetes outcomes earlier, but GLU results did identify additional outcomes. Moving from using only a coded diabetes diagnosis or medication dispensing to also using any GLU result values, preliminary diabetes outcomes identified increased an absolute 1.0% (17% relative increase).

At all sites the outcome rate was much higher among individuals with follow-up GLU results available, whether outcomes were determined using only coded diagnoses and medication dispensings or coded diagnoses, medication dispensings, and GLU laboratory results, suggesting bias in selecting individuals to receive glucose-related laboratory monitoring. This supports the possibility that clinicians selectively choose patients for glucose-related monitoring who they believe are at higher risk of diabetes.

For individuals with a baseline random or fasting glucose result value, there is no obvious relationship between the value of the baseline glucose result and missingness of a follow-up glucose result. However, for individuals with a baseline HbA1c result value, a higher baseline HbA1c result value appears to be a predictor of having a follow-up HbA1c result available.

Having a baseline GLU result available was strongly associated with having a follow-up GLU result available, but only at the large national insurer site. System and patient characteristics are potential contributors to this finding.

Olanzapine, quetiapine, and risperidone all had lower likelihood of missing follow-up GLU result values than did aripiprazole. We did not detect any difference in risk of diabetes within 365 days after SGA initiation by specific SGA.

### 2. Outcomes Detection Test Case 2: Upper gastrointestinal bleeding and hemoglobin (HGB) laboratory test results value outcomes among adults initiating a non-steroidal anti-inflammatory drug (NSAID)

#### a. Outcomes Detection Test Case 2 cohort development

Non-steroidal anti-inflammatory drugs (NSAIDs) are used for acute pain and for symptomatic treatment of chronic inflammatory and degenerative joint diseases. Their use is associated with a three- to five-fold increased risk of upper gastrointestinal (UGI) adverse events such as bleeding and perforated ulcer.[98] The use of selective cyclooxygenase (COX-2)-2 inhibitor NSAIDs is associated with a lower risk of UGI complications than is use of non-selective NSAIDs.[98] The intent of this outcomes test case was to determine whether including hemoglobin (HGB) result values obtained within 30 days after initiation of an NSAID increased the number of individuals identified with an UGI bleeding outcome.

Research questions addressed include:

1. What numbers (%) of patients have HGB results available? Specifically, within a cohort newly-starting an NSAID stratified by data partner site, describe the numbers and proportions of patients with HGB result values available before, after, and both before and after exposure to an NSAID from outpatient, inpatient, and emergency department care locations. (Note: Emergency department setting laboratory test results data are only available from one data partner site).
2. In the ambulatory, ambulatory to inpatient, inpatient, and emergency department to inpatient care locations, does use of HGB results identify additional cases of UGI bleeding beyond the use of diagnosis codes alone? Definitions include at least two HGB result values when one HGB is obtained after NSAID initiation and one is obtained either before NSAID exposure or prior to a bleeding outcome. UGI bleeding outcome is defined as a decrease of $\geq$ 3 g/dL between the HGB result values.
3. Independent of care location, does use of available HGB result values, either alone or in combination with outpatient diagnosis codes, identify additional cases of UGI bleeding beyond the use of inpatient diagnosis codes alone (i.e., the standard claims-based definition)?
4. For questions #1 through #3, describe UGI bleeding occurrence between individuals exposed to COX-2 selective vs. non-selective NSAIDs. NOTE: Sites 1 and 2 may have too low COX-2 use to analyze using this approach. If so, only non-selective NSAIDs will be used in analyses.
5. What is the confirmation rate of the diagnosis code (e.g., percentage of times change in HGB "confirms" the diagnosis)?

The exposures of interest included any NSAID newly-started (defined as not prescribed within the 183 days prior to the first NSAID dispensing in the date range) between January 1, 2008 and April 30, 2013 in individuals aged 18 years of age or older who had a minimum of 183 days enrollment in the health plan with medical and drug coverage prior to the first NSAID dispensing (bridging gaps up to 45 days). We selected individuals without a history of hematologic cancer, pregnancy, or diagnosed UGI bleeding during the 183 day baseline period (ICD-9 codes are in Appendix D). The cohort entry date ($T_0$) was the date of the first NSAID dispensing. Non-selective NSAIDs included diclofenac (oral and injection), etodolac, fenoprofen, flurbiprofen, ibuprofen (oral and injection), indomethacin (oral and rectal), ketoprofen, ketorolac (oral, injection, and nasal), meclofenamate, mefenamic acid, meloxicam, nabumetone, naproxen, oxaprozin, piroxicam, sulindac, tolmetin, prescription aspirin, choline magnesium salicylate, and diflunisal. The selective COX-2 Inhibitor was celecoxib.

The claims-based outcome definition was the presence of any single acute UGI bleeding or gastric ulcer diagnosis code (ICD-9 codes are in Appendix D) from an inpatient care location within 30 days after $T_0$. The claims plus laboratory results value-based (enriched) definition was the presence of any single UGI bleeding or gastric ulcer diagnosis code from a non-inpatient location AND a decrease of $\geq$ 3 g/dL between two HGB results (before and after $T_0$). This hierarchy was applied:

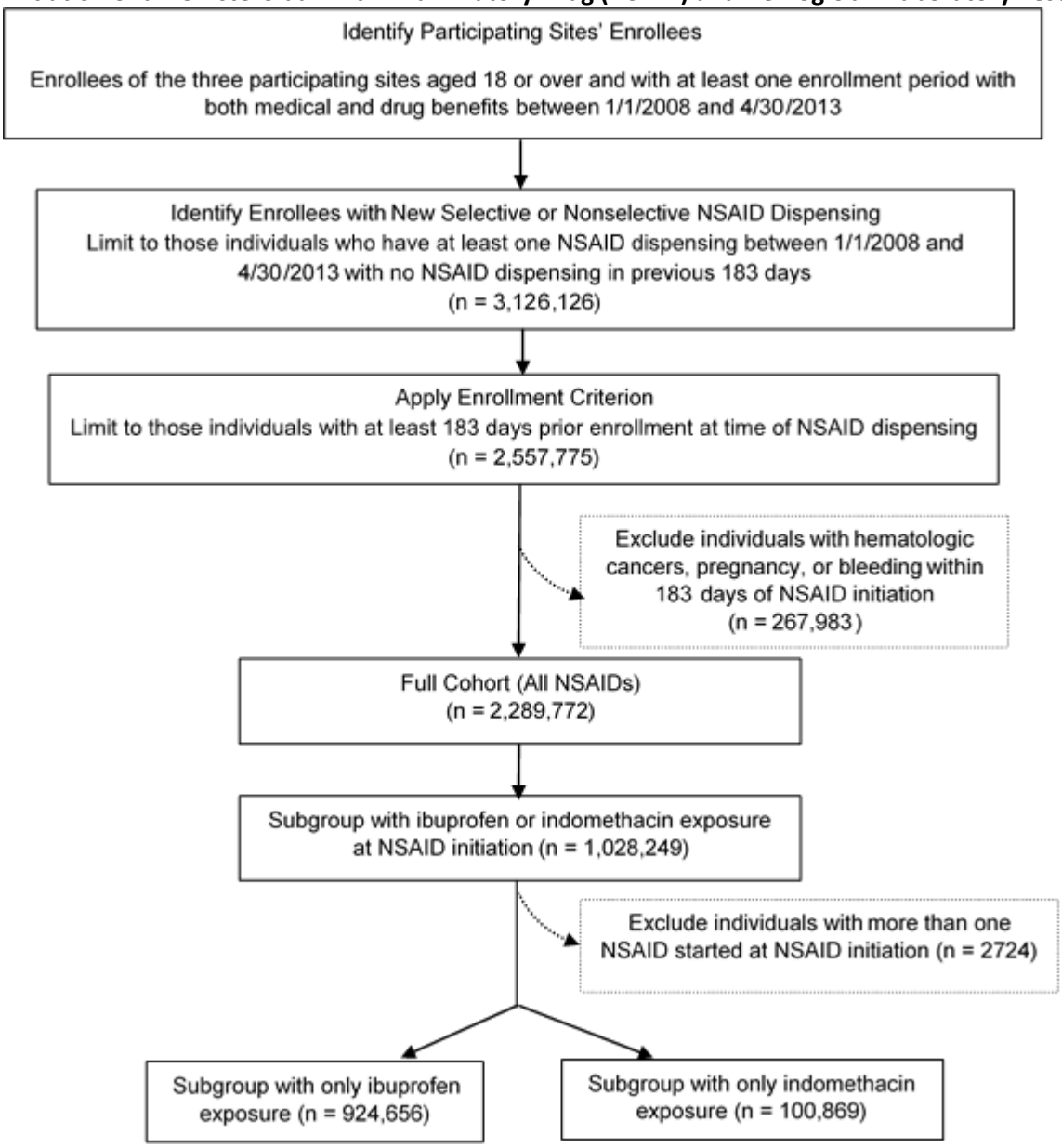
1. Identify patients with UGI bleeding or gastric ulcer outcome based on a coded inpatient diagnosis.
2. Among those who do not meet criterion #1, identify patients with a UGI bleeding or gastric ulcer outcome based on a coded non-inpatient diagnosis AND a drop in HGB of $\geq$ 3 g/dL.
3. Among those who do not meet criteria #1 or #2, identify patients with a drop in HGB of $\geq$ 3 g/dL (i.e., no coded bleeding diagnosis).

4. Among those who do not meet criteria #1, #2, or #3, identify patients with an UGI bleeding or gastric ulcer outcome based on a coded non-inpatient diagnosis who do not have a drop in HGB defined as either HGB results are available but the decrease is < 3 g/dL between two HGB results, or HGB results are not available.

While the workgroup intended to describe UGI bleeding for a selective and a non-selective NSAID, it was anticipated that celecoxib use could be very low at some sites. Therefore, the frequency of NSAID use across the three sites was first examined to determine the feasibility of the planned description. If not feasible, the workgroup would then decide the non-selective NSAIDs to examine based on the frequency of NSAID use.

Individuals in the cohort were censored at the first of death, discontinuation of medical or drug coverage, or end of study (October 31, 2013). For the bleeding outcomes descriptions, individuals were also censored when they were switched to an NSAID other than the NSAID dispensed on $T_0$. Covariates for this test case are shown in Appendix D.

**Figure 9. Test Case for Outcomes Detection Test Case 2, Gastrointestinal Bleeding among Adults after Initiation of a Non-steroidal Anti-inflammatory Drug (NSAID) and Hemoglobin Laboratory Test Results**



### b. Outcomes Detection Test Case 2 descriptive analysis of full cohort

***NSAID prescriptions in the full cohort***

Numbers of individuals dispensed each prescription NSAID among the 2,289,772 individuals from the three sites between January 1, 2008 and April 30, 2013 is shown in Table 46. Overall, the most common NSAID dispensed was ibuprofen (40.4%), followed by naproxen (15.5%), meloxicam (10.4%), diclofenac (8.0%), nabumetone (7.5%), and indomethacin (4.4%). Celecoxib accounted for 2.0% of new NSAID use, but 89.9% of the celecoxib prescriptions were from site 3 (40,871 of 45,449). Therefore, it was not appropriate to describe UGI bleeding occurrence for selective and non-selective NSAIDs for individual sites. The Workgroup decided to describe UGI bleeding occurrences for ibuprofen (n = 924,656) and indomethacin (n = 100,869).

**Table 46. New NSAID Exposures at Three Sites between January 1, 2008 and April 30, 2013, Overall and by Site**

| Non-Steroidal Anti-Inflammatory Drug | Numbers of Individuals with New Use | | | |
|---|---|---|---|---|
| | All Sites | Site 1 | Site 2 | Site 3 |
| Aspirin | 76,741 (3.4) | 1,617 (1.7) | 36,068 (3.2) | 39,056 (3.6) |
| Celecoxib | 45,449 (2.0) | 301 (0.3) | 4,277 (0.4) | 40,871 (3.8) |
| Choline magnesium salicylate | 1,318 (0.1) | 4 (0.0) | 1,013 (0.1) | 301 (0.0) |
| Diclofenac | 182,588 (8.0) | 16,930 (17.8) | 38,785 (3.5) | 126,873 (11.8) |
| Diflunisal | 2,484 (0.1) | 47 (0.0) | 114 (0.0) | 2,323 (0.2) |
| Etodolac | 67,460 (2.9) | 9,348 (9.8) | 32,604 (2.9) | 25,508 (2.4) |
| Fenoprofen | 485 (0.0) | 1 (0.0) | 5 (0.0) | 479 (0.0) |
| Flurbiprofen | 5,505 (0.2) | 1 (0.0) | 846 (0.1) | 4,658 (0.4) |
| Ibuprofen | 924,656 (40.4) | 24,974 (26.2) | 638,729 (57.0) | 260,953 (24.3) |
| Indomethacin | 100,869 (4.4) | 8400 (8.8) | 45,165 (4.0) | 47,304 (4.4) |
| Ketoprofen | 4,267 (0.2) | 459 (0.5) | 123 (0.0) | 3,685 (0.3) |
| Ketorolac | 66,511 (2.9) | 99 (0.1) | 3,978 (0.4) | 62,434 (5.8) |
| Meclofenamate | 863 (0.0) | 4 (0.0) | 665 (0.1) | 194 (0.0) |
| Mefenamic acid | 1,185 (0.1) | 0 (0.0) | 369 (0.0) | 816 (0.1) |
| Meloxicam | 238,728 (10.4) | 12,702 (13.3) | 30,235 (2.7) | 195,791 (18.2) |
| Nabumetone | 171,591 (7.5) | 5,959 (6.3) | 136,451 (12.2) | 29,181 (2.7) |
| Naproxen | 355,972 (15.5) | 12,484 (13.1) | 140,248 (12.5) | 203,240 (18.9) |
| Oxaprozin | 6,414 (0.3) | 5 (0.0) | 43 (0.0) | 6,366 (0.6) |
| Piroxicam | 9,998 (0.4) | 3 (0.0) | 680 (0.1) | 9,315 (0.9) |
| Salsalate | 6,576 (0.3) | 660 (0.7) | 3,419 (0.3) | 2,497 (0.2) |
| Sulindac | 12,845 (0.6) | 1,039 (1.1) | 4,760 (0.4) | 7,046 (0.7) |
| Tolmetin | 184 (0.0) | 29 (0.0) | 12 (0.0) | 143 (0.0 |
| $\geq$ 2 NSAIDs started the same date | 7,083 (0.3) | 116 (0.1) | 2,472 (0.2) | 4,495 (0.4) |

**Table 47** provides details of the full cohort overall and by site. Individuals starting an NSAID at site 3 were older (mean age 60.3 years) and had greater comorbidity. For example, across all three sites, 4.7% of individuals starting an NSAID had an anemia diagnosis, while at site 3, this percentage was 7.4% versus 1.8% and 2.4% at sites 1 and 2, respectively. Also, 16.2% of individuals from site 3 had a diagnosis of osteoarthritis, while this percentage was 8.6% and 5.7% at sites 1 and 2, respectively (**Table 47**). Within 30 days following NSAID initiation, 0.2% (N = 4991) had an UGI bleeding diagnosis (any care setting; 0.1% at sites 1 and 2, 0.3% at site 3).

**Table 47. Characteristics of Individuals in the Outcomes Detection Test Case 2 Population initiating an NSAID and Upper Gastrointestinal Bleeding Outcome within 30 Days by Site**

| Variable | All Sites (N = 2,289,772) | Site 1 (N = 95,182) | Site 2 (N = 1,121,061) | Site 3 (N = 1,073,529) |
|---|---|---|---|---|
| Upper gastrointestinal bleeding coded diagnosis within 30 days after cohort entry, N (%) yes | 4991 (0.2) | 109 (0.1) | 1325 (0.1) | 3557 (0.3) |
| Age in years, Mean (SD) | 54.1 ( 17.9 ) | 51.8 ( 16.7 ) | 48.4 ( 16.8 ) | 60.3 ( 17.1 ) |
| Gender, % female | 129,4241 (56.5) | 54,263 (57.0) | 621,155 (55.4) | 618,823 (57.6) |
| Race, N (%) | | | | |
| White | 1,230,305 (53.7) | 64,576 (67.8) | 624,145 (55.7) | 541,584 (50.4) |
| African American | 205,874 (9.0) | 5003 (5.3) | 105,461 (9.4) | 95,410 (8.9) |
| Other | 205,920 (9.0) | 3521 (3.7) | 185,249 (16.5) | 17,150 (1.6) |
| Unknown | 647,673 (28.3) | 22,082 (23.2) | 206,206 (18.4) | 419,385 (39.1) |
| Hispanic ethnicity, N (%) yes | 250,954 (11.0) | 13,598 (14.3) | 221,279 (19.7) | 16,077 (1.5) |
| Year of cohort entry | | | | |
| 2008 | 574,035 (25.1) | 22,364 (23.5) | 316,487 (28.2) | 235,184 (21.9) |
| 2009 | 476,178 (20.8) | 18,721 (19.7) | 244,394 (21.8) | 213,063 (19.8) |
| 2010 | 395,560 (17.3) | 17,007 (17.9) | 189,406 (16.9) | 189,147 (17.6) |
| 2011 | 367,272 (16.0) | 16,530 (17.4) | 166,774 (14.9) | 183,968 (17.1) |
| 2012 | 360,724 (15.8) | 15,543 (16.3) | 154,576 (13.8) | 190,605 (17.8) |
| 2013 | 116,003 (5.1) | 5017 (5.3) | 49,424 (4.4) | 61,562 (5.7) |
| Baseline hemoglobin CPT code, % yes [a] | 39.8 | 29.7 | 27.6 | 53.4 |
| Baseline hemoglobin result available, % yes [a] | 39.5 | 43.3 | 47.5 | 30.8 |
| Baseline hemoglobin value, mean (SD) [a] | 13.7 (1.6) | 14.5 (1.6) | 13.6 (1.6) | 13.6 (1.5) |
| Females mean (SD) | 13.1 (1.3) | 14.0 (1.4) | 13.0 (1.3) | 13.1 (1.3) |
| Males mean (SD) | 14.5 (1.5) | 15.5 (1.6) | 14.6 (1.5) | 14.4 (1.5) |
| Follow-up CPT code, % yes [b] | 17.2 | 15.3 | 14.6 | 20.1 |
| Follow-up hemoglobin result available, % yes [b] | 12.6 | 14.6 | 17.6 | 7.1 |
| Number of ambulatory medical visits during baseline, mean (SD) [a] | 5.4 ( 7.3 ) | 6.3 ( 8.4 ) | 4.0 ( 6.2 ) | 6.9 ( 8.0 ) |
| Emergency department visits during baseline, % yes [a] | 13.3 | 10.7 | 13.1 | 13.6 |
| Hospitalization during baseline, % yes [a] | 5.7 | 3.7 | 4.3 | 7.2 |
| Institutional stay during baseline, % yes [a] | 3.5 | 0.7 | 0.5 | 6.9 |
| No medical encounters during baseline, % with no encounters [a] | 14.5 | 12.5 | 19.6 | 9.4 |
| Number of unique therapeutic classes dispensed during baseline, mean (SD) [a] | 3.4 (3.7) | 2.8 (3.2) | 2.5 (3.1) | 4.5 (4.1) |
| Comorbidity score, mean (SD) [a] | 0.3 (1.3) | 0.3 (1.1) | 0.2 (1.0) | 0.4 (1.6) |
| Individual Comorbidities, N (%) yes [a] | | | | |
| Alcohol abuse | 23,135 (1.0) | 1332 (1.4) | 12,309 (1.1) | 9494 (0.9) |
| Anemia | 108,289 (4.7) | 1754 (1.8) | 27,003 (2.4) | 79,532 (7.4) |

| Variable | All Sites (N = 2,289,772) | Site 1 (N = 95,182) | Site 2 (N = 1,121,061) | Site 3 (N = 1,073,529) |
|---|---|---|---|---|
| Cardiac arrhythmias | 88,978 (3.9) | 2658 (2.8) | 21,948 (2.0) | 64,372 (6.0) |
| Chronic heart failure | 25,099 (1.1) | 554 (0.6) | 2329 (0.2) | 22,216 (2.1) |
| Coagulopathy | 14,727 (0.6) | 509 (0.5) | 3764 (0.3) | 10,454 (1.0) |
| Dementia | 10,969 (0.5) | 265 (0.3) | 1901 (0.2) | 8803 (0.8) |
| Fluid/Electrolyte disorders [c] | < 6 (0) | N/A | N/A | N/A |
| Complicated diabetes [c] | 90 (0.0) | N/A | N/A | N/A |
| Hemiplegia [c] | 7 (0) | N/A | N/A | N/A |
| AIDS/HIV | 3256 (0.1) | 42 (0.0) | 1473 (0.1) | 1741 (0.2) |
| Hypertension | 721,880 (31.5) | 18,241 (19.2) | 214,119 (19.1) | 489,520 (45.6) |
| Liver disease | 26,571 (1.2) | 1008 (1.1) | 12,202 (1.1) | 13,361 (1.2) |
| Metastatic cancer [c] | 9 (0) | N/A | N/A | N/A |
| Psychoses [c] | 205 (0.0) | N/A | N/A | N/A |
| Pulmonary circulation disorders [c] | 121 (0.0) | N/A | N/A | N/A |
| Chronic pulmonary disease | 94,559 (4.1) | 2174 (2.3) | 14,655 (1.3) | 77,730 (7.2) |
| Peripheral vascular disease | 57,922 (2.5) | 972 (1.0) | 10,699 (1.0) | 46,251 (4.3) |
| Renal failure | 11,008 (0.5) | 212 (0.2) | 2331 (0.2) | 8465 (0.8) |
| Any tumor | 33,268 (1.5) | 811 (0.9) | 10,569 (0.9) | 21,888 (2.0) |
| Weight loss | 929 (0.0) | 24 (0.0) | 305 (0.0) | 600 (0.1) |
| Osteoarthritis | 246,355 (10.8) | 8205 (8.6) | 63,900 (5.7) | 174,250 (16.2) |
| Rheumatoid arthritis | 28,990 (1.3) | 963 (1.0) | 6653 (0.6) | 21,374 (2.0) |
| Gastro-esophageal reflux | 185,528 (8.1) | 7765 (8.2) | 79,031 (7.0) | 98,732 (9.2) |
| Any cancer | 33,275 (1.5) | 811 (0.9) | 10,569 (0.9) | 21,895 (2.0) |
| Chronic kidney disease | 100,019 (4.4) | 3455 (3.6) | 30,664 (2.7) | 65,900 (6.1) |
| Selected Medications and Biologics [a] | | | | |
| Drugs that affect coagulation, N (%) yes | 46,338 (2.0) | 2085 (2.2) | 10795 (1.0) | 33,458 (3.1) |
| Misoprostol, N( %) yes | 2433 (0.1) | 68 (0.1) | 1581 (0.1) | 784 (0.1) |
| Prescription H2 blockers, N (%) yes | 109,264 (4.8) | 4513 (4.7) | 56,536 (5.0) | 48,215 (4.5) |
| Prescription proton pump inhibitors, N (%) yes | 254,245 (11.1) | 11,972 (12.6) | 80,975 (7.2) | 161,298 (15.0) |
| Antiplatelets, N (%) yes | 54,910 (2.4) | 858 (0.9) | 7813 (0.7) | 46,239 (4.3) |
| Oral glucocorticoids, N (%) yes | 162,711 (7.1) | 6254 (6.6) | 41,062 (3.7) | 115,395 (10.7) |
| Selective serotonin reuptake inhibitors, N (%) yes | 244,489 (10.7) | 11,816 (12.4) | 90,270 (8.1) | 142,403 (13.3) |
| Transfusion (packed cells or whole blood), N (%) yes | 1827 (0.1) | 50 (0.1) | 661 (0.1) | 1116 (0.1) |
| Epoetin, N (%) yes | 4791 (0.2) | 129 (0.1) | 1120 (0.1) | 3542 (0.3) |

[a] Within the 183 days prior to cohort entry
[b] Within the 30 days after cohort entry
[c] Site numbers not shown because some sites < 6

***Hemoglobin result value availability in the full cohort***

Patients with HGB results available before, after, and both before and after NSAID initiation are shown in Table 48. Within this cohort newly-starting an NSAID, 45.3% had at least one HGB result available within 365 days before or within 30 days after starting the NSAID (or both). HGB result availability varied from 33.8% at site 3, to 50.4% at site 1, and to 55.8% at site 2. Overall, 32.7% had a HGB result available only before, 5.8% had a HGB result available only within 30 days after, and 6.8% had HGB results available both before and after NSAID initiation.

HGB result availability by care location where the laboratory specimen was obtained is also shown in Table 48. Only site 2 had laboratory results from the emergency department. Some patients with HGB results only within the 365 days before NSAID initiation had results from more than one care location (i.e., total > 100%), but at all sites over 90% of patients with HGB results obtained only before NSAID initiation had HGB results from the outpatient setting. In addition, at sites 1 and 2 respectively, 13.4% and 7.8% of patients with HGB results only before NSAID initiation had results from inpatient settings. At site 2, 19.6% of patients with HGB results obtained only before NSAID initiation had HGB results from the emergency department. Similar patterns of care locations where HGB results were obtained were observed for patients with results only after NSAID initiation and for patients with HGB results both before and after NSAID initiation (Table 48). Among patients with HGB results available both before and after NSAID initiation, 76.1% had results from the outpatient setting, but this proportion was heavily influenced by 99.0% of the HGB results being from the outpatient location at site 3. For sites 1 and 2, higher proportions of HGB results were obtained from non-outpatient locations. ***This highlights the importance of missingness of laboratory result values from the inpatient and emergency department care settings when an acute outcome is being assessed.***

**Table 48. Hemoglobin Laboratory Result Value Availability within 365 days before and within 30 days after starting an NSAID, Overall and by Site**

| Hemoglobin Result Value Availability [a, b, c] | Overall (N = 2,289,772) | Site 1 (N = 95,182) | Site 2 (N = 1,121,061) | Site 3 (N = 1,073,529) |
|---|---|---|---|---|
| Available at any time within 365 days before to 30 days after NSAID Initiation | | | | |
| Yes | 1,036,29419 (45.3) | 47,947 (50.4) | 625,961 (55.8) | 363,075 (33.8) |
| No | 1,253,553 (54.7) | 47,235 (49.6) | 495,864 (44.2) | 710,454 (66.2) |
| Timing of Availability | | | | |
| Before NSAID Initiation Only | 747,640 (32.7) | 34,034 (35.8) | 426,798 (38.1) | 286,808 (26.7) |
| After NSAID Initiation Only | 132,507 (5.8) | 6773 (7.1) | 92,987 (8.3) | 32,747 (3.1) |
| Both before and after NSAID initiation | 156,072 (6.8) | 7140 (7.5) | 105,412 (9.4) | 43,520 (4.1) |
| | | | | |
| **Care Setting where Laboratory Specimen was Obtained** | | | | |
| **Before NSAID initiation only [a, b, c]** | **Overall (N =747,640)** | **Site 1 (N = 34,034)** | **Site 2 (N = 426,798)** | **Site 3 (N = 286,808)** |
| Emergency department | 83,503 (11.2) | N/A | 83,503 (19.6) | N/A |
| Inpatient | 40,320 (5.4) | 4548 (13.4) | 33,407 (7.8) | 2365 (0.8) |
| Outpatient | 707,746 (94.7) | 32,661 (96.0) | 389,710 (91.3) | 285,375 (99.5) |
| | | | | |
| **After NSAID initiation only [a, b, c]** | **Overall (N = 132,507)** | **Site 1 (N = 6773)** | **Site 2 (N = 92,987)** | **Site 3 (N =32,747)** |
| Emergency department | 17,007 (12.8) | N/A | 17,007 (18.3) | N/A |
| Inpatient | 3069 (2.3) | 470 (6.9) | 2276 (2.4) | 323 (1.0) |
| Outpatient | 117,434 (88.6) | 6442 (95.1) | 78,537 (84.5) | 32,455 (99.1) |
| | | | | |
| **Both before and after NSAID initiation [a, b, c]** | **Overall (N = 156,072)** | **Site 1 (N = 7140)** | **Site 2 (N = 105,412)** | **Site 3 (N = 43,520)** |
| Emergency and Emergency | 12,344 (7.9) | N/A | 12,344 (11.7) | N/A |
| Emergency and Inpatient | 12,650 (8.1) | N/A | 12,650 (12.0) | N/A |
| Emergency and Outpatient | 21,957 (14.1) | N/A | 21,957 (20.8) | N/A |
| Inpatient and Emergency | 6315 (4.0) | N/A | 6315 (6.0) | N/A |
| Inpatient and Inpatient | 14,016 (9.0) | 524 (7.3) | 13,227 (12.5) | 265 (0.6) |
| Inpatient and Outpatient | 14,217 (9.1) | 1177 (16.5) | 12,768 (12.1) | 272 (0.6) |
| Outpatient and Emergency | 22,056 (14.1) | N/A | 22,056 (20.9) | N/A |
| Outpatient and Inpatient | 18,123 (11.6) | 1030 (14.4) | 16,820 (16.0) | 273 (0.6) |
| Outpatient and Outpatient | 118,695 (76.1) | 6135 (85.9) | 69,481 (65.9) | 43,079 (99.0) |

[a] Only Site 2 has emergency department laboratory test results available in the MSDD

[b] Outpatient setting reflects outpatient and "unknown" locations considered together because Mini-Sentinel Data Partners have stated that laboratory results with the setting variable populated with "unknown" are primarily outpatient laboratory test results

[c] The N in each setting do not add to the N for the "Any" setting or "Overall" numbers because the same individual could have had HGB results from more than one setting (i.e., same individual could be counted in different settings)

***Upper gastrointestinal bleeding outcomes in the full cohort***

As shown in Table 49, inpatient diagnoses (regardless of whether or not HGB results were available to document a HGB drop ≥ 3 g/dL; group 1), accounted for 1657 (21.7%) of the UGI bleeding cases across the sites (details of specific coded diagnoses in Appendix I). Using HGB results indicating a HGB drop ≥ 3 g/dL together with an UGI bleeding diagnosis assigned in a non-inpatient setting (group 2) only identified an additional 58 cases (0.8%) across all sites. Using a HGB drop ≥ 3 g/dL and no coded UGI bleeding diagnosis (group 3) added 2619 (34.3%) cases (Appendix I). Non-inpatient diagnoses without available HGB results or with a HGB drop < 3 g/dL accounted for 3303 (43.4%) cases (group 4). Care locations where HGB results were obtained (for identifying HGB drops ≥ 3 g/dL in groups 2 and 3) are in Appendix I.

**Table 49. Upper Gastrointestinal (UGI) Bleeding Outcomes within 30 Days in the Full Cohort after NSAID Initiation using Varied Outcomes Definitions, Overall and by Site**

| Group | Bleeding Outcomes Criteria (mutually exclusive groups) | Data Partner Site | | | |
|---|---|---|---|---|---|
| | | All Sites | Site 1 | Site 2 | Site 3 |
| 1 | Inpatient diagnoses (with or without an observed HGB drop ≥ 3 g/dL) | 1657 (21.7) | 30 (11.7) | 520 (14.9) | 1107 (28.4) |
| 2 | Non-inpatient diagnosis with drop in HGB ≥ 3 g/dL | 58 (0.8) | 2 (0.8) | 41 (1.2) | 15 (0.4) |
| 3 | Observed drop in HBG ≥ 3 g/dL (no coded UGI bleeding diagnosis) | 2619 (34.3) | 148 (57.6) | 2160 (61.9) | 311 (8.0) |
| 4 | Non-inpatient diagnosis without observed drop in HGB | 3303 (43.3) | 77 (30.0) | 769 (22.0) | 2457 (63.2) |
| 1 – 4 | Total bleeding outcomes [a] | 7637 | 257 | 3490 | 3890 |
| **Group** | **Bleeding Outcomes Criteria Excluding Group 3 (mutually exclusive groups)** | **All Sites** | **Site 1** | **Site 2** | **Site 3** |
| 1 | Inpatient diagnoses (with or without an observed HGB drop ≥ 3 g/dL) | 1657 (33.0) | 30 (27.5) | 520 (39.1) | 1107 (30.9) |
| 2 | Non-inpatient diagnosis with drop in HGB ≥ 3 g/dL | 58 (1.2) | 2 (1.8) | 41 (3.1) | 15 (0.4) |
| 4 | Non-inpatient diagnosis without observed drop in HGB | 3303 (65.8) | 77 (70.6) | 769 (57.8) | 2457 (68.7) |
| 1, 2, 4 | Total UGI bleeding outcomes without Group 3 [a] | 5018 | 109 | 1330 | 3579 |

[a] Includes all UGI bleeding outcomes within 30 days after $T_0$, whether or not the individual had an NSAID switch within those 30 days. Therefore, the total N in this row differs from the total N of UGI bleeding outcomes shown in Table 47 because it also includes UGI bleeding outcomes in the 27 additional individuals (site 2 = 5, site 3 = 22) who had an UGI bleed within 30 days after $T_0$ but after the date of switching to an NSAID other than the NSAID dispensed on $T_0$.

Importantly, because some patients in group 3 were anticipated to have a coded bleeding diagnosis from a non-UGI site, to further understand the cases identified using only a drop in HGB ≥ 3 g/dL, the Workgroup examined all coded bleeding diagnoses in group 3 patients. We found only 250 (9.5%) of group 3 patients had any coded non-UGI bleeding diagnosis (Appendix I). Among these 250 patients, common sites included obstetric or gynecologic, pulmonary, and intracranial (subarachnoid, cerebellar, intracerebral, subdural). To further characterize the patients captured in group 3, the Workgroup investigated the overall array of diagnoses occurring between NSAID initiation and bleeding outcome within these patients. Across all sites, the most frequently coded diagnoses included high-prevalence conditions (e.g., hypertension, dyslipidemia, diabetes), the use of therapies associated with potential bleeding or the consequences of bleeding (e.g., long-term aspirin, anticoagulant therapy, acute post-hemorrhagic anemia), non-hemorrhagic conditions associated with anemia (e.g., autoimmune and inflammatory diseases, chronic infections, chronic conditions), and indications of use (e.g., pain). In light of these findings, the Workgroup does not recommend the use of a HGB drop ≥ 3 g/dL and no coded UGI bleeding diagnosis (group 3) to identify UGI bleeding or bleeding in general, due to the underperformance in identifying specific bleeding sites or bleeding events in general versus non-hemorrhagic conditions.

If the 2619 group 3 cases are not included as part of the total UGI bleeding cases within 30 days after NSAID initiation (bottom section of Table 49), the number of UGI bleeding cases drops more dramatically at sites 1 and 2 than at site 3. As a result, the percentage of cases contributed by inpatient diagnoses (group 1) and by non-inpatient diagnoses without an observed drop in HGB (group 4) shifts to a greater degree at sites 1 and 2 than at site 3. ***Availability of HGB results alters estimated numbers (and rates) of UGI bleeding outcomes after NSAID initiation more at sites that have a higher percentage of the population with laboratory results available than at sites where laboratory results are missing in a higher percentage of the population.***

Because the definition used in group 4 (non-inpatient diagnoses without available HGB results or with a HGB drop < 3 g/dL) identified a much larger proportion of UGI bleeding events compared with group 1 and 2, the Workgroup examined whether the distribution of the most frequently coded UGI diagnoses in group 4 was similar to that observed in groups 1 and 2. In general, there was reasonably good agreement between specific UGI diagnoses identified in group 4 and other groups. As an example, across all sites the most frequent five coded UGI diagnoses in group 4 (UGI hemorrhage, peptic ulcer, antral ulcer, hematemesis, and duodenal ulcer) overlapped with the four most frequently occurring UGI diagnoses in groups 1 and 2 (Appendix I, Tables 1-3). Good agreement was also found within each site when individually assessed. On the basis of these findings, the Workgroup agrees that the definition used to identify bleeding events in group 4 may be a reasonable strategy to consider for UGI bleeding definition augmentation.

***Confirmation rate of upper gastrointestinal bleeding diagnosis code in the full cohort: proportions of available HGB result values consistent with HGB decrease ≥ 3 g/dL***

As shown in Table 50, among patients with bleeding identified through a coded inpatient diagnosis (group 1), 13.9% could be confirmed as having a drop in HGB ≥ 3 g/dL using HGB result values available in the MSDD before and after NSAID initiation, 15.6% had a drop in HGB < 3 g/dL confirmed using HGB result values available before and after NSAID initiation, and 70.5% had no HGB results data available before and/or after NSAID initiation (i.e., the bleed could not be confirmed using available laboratory results). Among the patients with bleeding identified through a coded diagnosis from the non-inpatient setting (groups 2 and 4), 1.7% were confirmed as having a drop in HGB ≥ 3 g/dL using HGB result values

available before and after NSAID initiation (the 58 individuals in group 2), 12.2% had a drop in HGB < 3 g/dL confirmed using two HGB result values available before and after NSAID initiation, and 80.9% had no HGB results available before and/or after NSAID initiation (group 4).

Table 50 also provides information about numbers and percentages of individuals in each group who had HGB laboratory procedure claims codes (CPT). For example, in group 1, of the 230 (13.9%) individuals with a HGB value drop ≥ 3 g/dL, 81.3% had a HGB CPT code both before and after NSAID initiation; of the 1168 (70.5%) individuals with HGB results missing, 30.2% had a HGB CPT code both before and after NSAID initiation. This disparity between HGB result values and CPT claims further confirms the missingness of HGB laboratory results in the MSDD.

**Table 50. Confirmation of Upper Gastrointestinal Bleeding Diagnosis Code in the Full Cohort by Available Hemoglobin Result Values Consistent with HGB Decrease > 3 g/dL**

| Group | HGB Value Decrease ≥ 3 g/dL [d] | | | | HGB Value Decrease < 3 g/dL | | | | HGB results missing before and/or after NSAID initiation (i.e., unable to determine HGB change) | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CPT both [a] | CPT one [b] | CPT neither [c] | Subtotal | CPT both | CPT one | CPT neither | Subtotal | CPT both | CPT one | CPT neither | Subtotal | |
| 1 (inpatient diagnosis)(N = 1657) | 187 (81.3) | 42 (18.3) | 1 (0.4) | 230 (13.9) | 215 (83.0) | 41 (15.8) | 3 (1.2) | 259 (15.6) | 353 (30.2) | 644 (55.1) | 171 (14.6) | 1168 (70.5) | 1657 |
| 2 (non-inpatient diagnosis with drop in HGB > 3 g/dL)(N = 58) | 49 (84.5) | 8 (13.8) | 1 (1.7) | 58 (1.7) | | | | | | | | | |
| 4 (non-inpatient diagnosis without observed drop in HGB)(N = 3303) | | | | | 331 (80.9) | 68 (16.6) | 10 (2.4) | 409 (12.2) | 728 (25.2) | 1606 (55.5) | 560 (19.4) | 2894 (86.1) | 3361 |
| 1, 2, and 4 Total | 236 (81.9) | 50 (17.4) | 2 (0.7) | 288 (5.7) | 546 (81.7) | 109 (16.3) | 13 (1.9) | 668 (13.3) | 1081 (26.6) | 2250 (55.4) | 731 (18.0) | 4062 (80.9) | 5018 |

[a] CPT code for HGB test before and after NSAID initiation
[b] CPT code for HGB test before or after NSAID initiation but not both
[c] No CPT code for HGB test
[d] Percentages in CPT columns are row percentages of the corresponding Subtotal column; percentages in Subtotal columns are row percentages of Total column

### c. Outcomes Detection Test Case 2 descriptive analysis of ibuprofen and indomethacin sub-cohort

***NSAID prescriptions in the ibuprofen and indomethacin sub-cohort***

Table 51 provides details of the sub-cohort starting ibuprofen (N = 924,656) or indomethacin (N = 100,869). Individuals starting ibuprofen were younger (mean age 47.1 years) and included a higher proportion of women (57.0%) than individuals starting indomethacin (mean age 58.1 years; 30.8% women). In general, patients starting ibuprofen or indomethacin from site 3 had greater comorbidity. For example, across all three sites, less than 1% (0.7%) of individuals starting ibuprofen took an anticoagulant, and 3.2% of individuals starting indomethacin took an anticoagulant. However, at site 3 those percentages were 1.7% and 5.3%, respectively (Table 51).

Within 30 days following ibuprofen initiation, 0.1% (N = 1066) had an UGI bleeding diagnosis (any care setting; 0.1% at sites 1 and 2, 0.2% at site 3). Within 30 days after indomethacin initiation, 0.4% (N = 378) of individuals had an UGI bleeding diagnosis (any care setting; 0.1% at site 1, 0.2% at site 2, 0.6% at site 3).

**Table 51. Characteristics of Individuals in the Outcomes Detection Test Case 2 Population, Initiation of Ibuprofen or Indomethacin and Upper Gastrointestinal Bleeding Outcome within 30 Days by Site**

| Variable | Ibuprofen | | | | Indomethacin | | | |
|---|---|---|---|---|---|---|---|---|
| | Site 1 (N = 24,974) | Site 2 (N = 638,729) | Site 3 (N = 260,953) | All Sites (N = 924,656) | Site 1 (N = 8400) | Site 2 (N = 45,165) | Site 3 (N = 47,304) | All Sites (N = 100,869) |
| Upper gastrointestinal bleeding coded diagnosis within 30 days after cohort entry, N (%) yes | 22 (0.1) | 428 (0.1) | 616 (0.2) | 1066 (0.1) | 12 (0.1) | 91 (0.2) | 275 (0.6) | 378 (0.4) |
| Age in years, Mean (SD) | 46.1 (14.9) | 44.3 (15.3) | 54.1 (18.1) | 47.1 (16.7) | 50.7 (13.9) | 53.5 (15.3) | 63.8 (14.5) | 58.1 (15.8) |
| Gender, % female | 15,863 (63.5) | 359,585 (56.3) | 151,222 (57.9) | 526,670 (57.0) | 1994 (23.7) | 13,635 (30.2) | 15,401 (32.6) | 31,030 (30.8) |
| Race, N (%) | | | | | | | | |
| White | 14,987 (60.0) | 332,829 (52.1) | 97,185 (37.2) | 445,001 (48.1) | 5799 (69.0) | 24,873 (55.1) | 26,675 (56.4) | 57,347 (56.9) |
| African American | 2037 (8.2) | 67,485 (10.6) | 21,915 (8.4) | 91,437 (9.9) | 388 (4.6) | 3285 (7.3) | 5313 (11.2) | 8986 (8.9) |
| Other | 1043 (4.2) | 106,260 (16.6) | 3916 (1.5) | 111,219 (12.0) | 404 (4.8) | 10,763 (23.8) | 957 (2.0) | 12,124 (12.0) |
| Unknown | 6907 (27.7) | 132,155 (20.7) | 137,937 (52.9) | 276,999 (30.0) | 1809 (21.5) | 6244 (13.8) | 14,359 (30.4) | 22,412 (22.2) |
| Hispanic ethnicity, N (%) yes | 4564 (18.3) | 138,533 (21.7) | 4141 (1.6) | 147,238 (15.9) | 887 (10.6) | 5815 (12.9) | 358 (0.8) | 7060 (7.0) |
| Year of cohort entry | | | | | | | | |
| 2008 | 6151 (24.6) | 183,914 (28.8) | 58,119 (22.3) | 248,184 (26.8) | 2319 (27.6) | 14,916 (33.0) | 11,663 (24.7) | 28,898 (28.6) |
| 2009 | 5100 (20.4) | 139,556 (21.8) | 52,167 (20.0) | 196,823 (21.3) | 1800 (21.4) | 10,778 (23.9) | 9802 (20.7) | 22,380 (22.2) |
| 2010 | 4387 (17.6) | 105,806 (16.6) | 45,895 (17.6) | 156,088 (16.9) | 1507 (17.9) | 7240 (16.0) | 8079 (17.1) | 16,826 (16.7) |
| 2011 | 4234 (17.0) | 93,537 (14.6) | 43,468 (16.7) | 141,239 (15.3) | 1304 (15.5) | 5819 (12.9) | 7883 (16.7) | 15,006 (14.9) |
| 2012 | 3790 (15.2) | 8,7823 (13.7) | 46,136 (17.7) | 137,749 (14.9) | 1148 (13.7) | 5034 (11.1) | 7648 (16.2) | 1,830 (13.7) |

| Variable | Ibuprofen | | | | Indomethacin | | | |
|---|---|---|---|---|---|---|---|---|
| | Site 1 (N = 24,974) | Site 2 (N = 638,729) | Site 3 (N = 260,953) | All Sites (N = 924,656) | Site 1 (N = 8400) | Site 2 (N = 45,165) | Site 3 (N = 47,304) | All Sites (N = 100,869) |
| 2013 | 1312 (5.3) | 28,093 (4.4) | 15,168 (5.8) | 44573 (4.8) | 322 (3.8) | 1378 (3.1) | 2229 (4.7) | 3929 (3.9) |
| Baseline hemoglobin CPT code, % yes [a] | 30.0 | 24.1 | 47.8 | 31.0 | 25.9 | 29.3 | 54.5 | 40.8 |
| Baseline hemoglobin result available, % yes [a] | 41.6 | 43.3 | 29.5 | 39.3 | 37.9 | 49.7 | 27.2 | 38.2 |
| Baseline hemoglobin value, mean (SD) [a] | 14.4 (1.7) | 13.7 (1.6) | 13.6 (1.5) | 13.7 (1.6) | 15.2 (1.6) | 14.0 (1.8) | 13.8 (1.6) | 14.0 (1.7) |
| Females mean (SD) | 13.9 (1.5) | 13.0 (1.3) | 13.1 (1.3) | 13.1 (1.3) | 14.1 (1.4) | 13.0 (1.5) | 13.0 (1.3) | 13.1 (1.5) |
| Males mean (SD) | 15.7 (1.5) | 14.8 (1.3) | 14.5 (1.4) | 14.8 (1.4) | 15.8 (1.5) | 14.6 (1.6) | 14.3 (1.6) | 14.6 (1.7) |
| Follow-up CPT code, % yes [b] | 15.0 | 13.1 | 17.5 | 14.4 | 22.0 | 20.9 | 27.6 | 24.1 |
| Follow-up hemoglobin result available, % yes [b] | 14.1 | 15.8 | 16.1 | 13.0 | 21.1 | 24.1 | 9.0 | 16.7 |
| Number of ambulatory medical visits during baseline, mean (SD) [a] | 4.8 (6.8) | 3.2 (5.2) | 5.5 (7.2) | 3.8 (5.9) | 4.9 (6.9) | 4.3 (6.9) | 6.6 (7.6) | 5.4 (7.3) |
| Emergency department visits during baseline, % yes [a] | 11.7 | 12.0 | 15.1 | 12.9 | 9.5 | 12.3 | 14.0 | 12.8 |
| Hospitalization during baseline, % yes [a] | 4.1 | 2.5 | 6.2 | 3.6 | 2.8 | 4.4 | 7.7 | 5.8 |
| Institutional stay during baseline, % yes [a] | 0.2 | 0.2 | 6.0 | 1.8 | 0.2 | 2.0 | 6.7 | 4.1 |
| No medical encounters during baseline, % with no encounters [a] | 14.6 | 22.6 | 14.4 | 20.1 | 19.1 | 21.3 | 10.2 | 15.9 |
| Number of unique therapeutic classes dispensed during baseline, mean (SD) [a] | 2.2 (2.7) | 2.0 (2.6) | 3.8 (3.8) | 2.5 (3.1) | 2.5 (2.9) | 3.1 (3.3) | 4.9 (4.1) | 3.9 (3.8) |
| Comorbidity score, mean (SD) [a] | 0.2 (0.8) | 0.1 (0.7) | 0.3 (1.4) | 0.2 (0.9) | 0.1 (0.9) | 0.2 (1.4) | 0.5 (1.7) | 0.3 (1.5) |
| Individual Comorbidities, N (%) yes [a] | | | | | | | | |
| Alcohol abuse | 368 (1.5) | 6178 (1.0) | 2592 (1.0) | 9138 (1.0) | 155 (1.8) | 703 (1.6) | 550 (1.2) | 1408 (1.4) |
| Anemia | 481 (1.9) | 11173 (1.7) | 16702 (6.4) | 28356 (3.1) | 72 (0.9) | 1269 (2.8) | 3489 (7.4) | 4830 (4.8) |
| Cardiac arrhythmias | 332 (1.3) | 5049 (0.8) | 10355 (4.0) | 15736 (1.7) | 163 (1.9) | 1140 (2.5) | 4044 (8.5) | 5347 (5.3) |
| Chronic heart failure | 77 (0.3) | 508 (0.1) | 3814 (1.5) | 4399 (0.5) | 36 (0.4) | 125 (0.3) | 1372 (2.9) | 1533 (1.5) |
| Coagulopathy | 78 (0.3) | 970 (0.2) | 2027 (0.8) | 3075 (0.3) | 40 (0.5) | 209 (0.5) | 597 (1.3) | 846 (0.8) |
| Dementia [c] | 18 (0.1) | 415 (0.1) | 1612 (0.6) | 2045 (0.2) | N/A | N/A | N/A | 439 (0.4) |
| Complicated diabetes [c] | N/A | N/A | N/A | 15 (0.0) | N/A | N/A | N/A | < 6 (0.0) |
| Fluid/Electrolyte dis orders [c] | N/A | N/A | N/A | < 6 (0.0) | N/A | N/A | N/A | < 6 (0.0) |
| Hemiplegia [c] | N/A | N/A | N/A | < 6 (0.0) | N/A | N/A | N/A | < 6 (0.0) |
| AIDS/HIV | 15 (0.1) | 868 (0.1) | 637 (0.2) | 1520 (0.2) | 7 (0.1) | 74 (0.2) | 51 (0.1) | 132 (0.1) |

| Variable | Ibuprofen | | | | Indomethacin | | | |
|---|---|---|---|---|---|---|---|---|
| | Site 1 (N = 24,974) | Site 2 (N = 638,729) | Site 3 (N = 260,953) | All Sites (N = 924,656) | Site 1 (N = 8400) | Site 2 (N = 45,165) | Site 3 (N = 47,304) | All Sites (N = 100,869) |
| Hypertension | 3296 (13.2) | 86,818 (13.6) | 94,598 (36.3) | 184,712 (20.0) | 1768 (21.0) | 12,409 (27.5) | 25,975 (54.9) | 40,152 (39.8) |
| Liver disease | 276 (1.1) | 6093 (1.0) | 3598 (1.4) | 9967 (1.1) | 91 (1.1) | 569 (1.3) | 580 (1.2) | 1240 (1.2) |
| Metastatic cancer [c] | N/A | N/A | N/A | < 6 (0.0) | N/A | N/A | N/A | < 6 (0.0) |
| Psychoses [c] | N/A | N/A | N/A | 60 (0.0) | N/A | N/A | N/A | 6 (0.0) |
| Pulmonary circulation disorder [c] | N/A | N/A | N/A | 22 (0.0) | N/A | N/A | N/A | 6 (0.0) |
| Chronic pulmonary disease | 270 (1.1) | 4338 (0.7) | 14,295 (5.5) | 18,903 (2.0) | 116 (1.4) | 885 (2.0) | 3800 (8.0) | 4801 (4.8) |
| Peripheral vascular disease | 103 (0.4) | 2484 (0.4) | 7977 (3.1) | 10,564 (1.1) | 71 (0.8) | 621 (1.4) | 2520 (5.3) | 3212 (3.2) |
| Renal failure | 17 (0.1) | 372 (0.1) | 1396 (0.5) | 1785 (0.2) | 7 (0.1) | 124 (0.3) | 695 (1.5) | 826 (0.8) |
| Any tumor | 139 (0.6) | 4346 (0.7) | 4461 (1.7) | 8946 (1.0) | 70 (0.8) | 716 (1.6) | 1507 (3.2) | 2293 (2.3) |
| Weight loss [c] | N/A | N/A | N/A | 170 (0.0) | N/A | N/A | N/A | 77 (0.1) |
| Osteoarthritis | 1246 (5.0) | 20,712 (3.2) | 26,114 (10.0) | 48,072 (5.2) | 458 (5.5) | 2547 (5.6) | 6222 (13.2) | 9227 (9.1) |
| Rheumatoid arthritis | 163 (0.7) | 2191 (0.3) | 3350 (1.3) | 5704 (0.6) | 34 (0.4) | 248 (0.5) | 741 (1.6) | 1023 (1.0) |
| Gastro-esophageal reflux | 1672 (6.7) | 32,535 (5.1) | 19,884 (7.6) | 54,091 (5.8) | 539 (6.4) | 3076 (6.8) | 3630 (7.7) | 7245 (7.2) |
| Any cancer | 139 (0.6) | 4346 (0.7) | 4461 (1.7) | 8946 (1.0) | 70 (0.8) | 716 (1.6) | 1507 (3.2) | 2293 (2.3) |
| Chronic kidney disease | 298 (1.2) | 7169 (1.1) | 11689 (4.5) | 19156 (2.1) | 215 (2.6) | 2265 (5.0) | 4471 (9.5) | 6951 (6.9) |
| Selected Medications and Biologics [a] | | | | | | | | |
| Drugs that affect coagulation, N (%) yes | 204 (0.8) | 2182 (0.3) | 4485 (1.7) | 6871 (0.7) | 102 (1.2) | 619 (1.4) | 2525 (5.3) | 3246 (3.2) |
| Misoprostol, N( %) yes | 41 (0.2) | 1256 (0.2) | 329 (0.1) | 1626 (0.2) | N/A | N/A | N/A | 37 (0.0) |
| Prescription H2 blockers, N (%) yes | 933 (3.7) | 25,596 (4.0) | 9475 (3.6) | 36,004 (3.9) | 346 (4.1) | 2571 (5.7) | 2344 (5.0) | 5261 (5.2) |
| Prescription proton pump inhibitors, N (%) yes | 2402 (9.6) | 33,715 (5.3) | 30,347 (11.6) | 66,464 (7.2) | 835 (9.9) | 3227 (7.1) | 6231 (13.2) | 10,293 (10.2) |
| Antiplatelets, N (%) yes | 87 (0.3) | 2144 (0.3) | 7129 (2.7) | 9360 (1.0) | 83 (1.0) | 490 (1.1) | 2924 (6.2) | 3497 (3.5) |
| Oral glucocorticoids, N (%) yes | 1377 (5.5) | 18,810 (2.9) | 22,480 (8.6) | 42,667 (4.6) | 682 (11.9) | 2557 (5.7) | 6022 (12.7) | 15,164 (9.2) |
| Selective serotonin reuptake inhibitors, N (%) yes | 2898 (11.6) | 45,646 (7.1) | 31,044 (11.9) | 79,588 (8.6) | 635 (7.6) | 3057 (6.8) | 4481 (9.5) | 8173 (8.1) |
| Transfusion (packed cells or whole blood), N (%) yes | 25 (0.1) | 276 (0.0) | 271 (0.1) | 572 (0.1) | N/A | N/A | N/A | 173 (0.2) |
| Epoetin, N (%) yes | N/A | N/A | N/A | 855 (0.1) | N/A | N/A | N/A | 345 (0.3) |

[a] Within the 183 days prior to cohort entry
[b] Within the 30 days after cohort entry
[c] Individual cell numbers not shown because some N = 6

***Hemoglobin result value availability in the ibuprofen and indomethacin sub-cohort***

Numbers and percentages of patients with HGB results available before, after, and both before and after ibuprofen or indomethacin was started are shown in **Table 52**. Across this sub-cohort, 53.2% of individuals starting ibuprofen and 52.9% of individuals starting indomethacin had at least one HGB result available within 365 days before or within 30 days after the NSAID was started (or both). HGB result availability varied more among patients starting indomethacin, ranging from 36.2% at site 3 to 69.6% at site 2. Overall, among individuals starting indomethacin, 38.2% had a HGB result available only before, 16.8% had a HGB result available only within 30 days after, and 8.3% had HGB results available both before and after initiation. Overall, among individuals starting ibuprofen, 39.3% had a HGB result available only before, 13.3% had had a HGB result available only after, and 6.4% had HGB results available both before and after initiation. Site-specific information for both NSAIDs is also in **Table 52**.

HGB result availability by care location where the laboratory specimen was obtained is also shown in **Table 52**. Patterns of HGB result availability by care location for this sub-cohort of individuals are similar the full cohort, once again highlighting the importance of missing laboratory result values from the inpatient and emergency department settings when an acute outcome is being assessed.

**Table 52. Hemoglobin Laboratory Result Value Availability within 365 days before and within 30 days after Starting Ibuprofen or Indomethacin, Overall and by Site**

| Hemoglobin Result Value Availability Before and After Ibuprofen or Indomethacin Initiation [a, b, c, d] | Ibuprofen | | | | Indomethacin | | | |
|---|---|---|---|---|---|---|---|---|
| | **All Sites** (N = 924,656) | **Site 1** (N = 24,974) | **Site 2** (N = 638,729) | **Site 3** (N = 260,953) | **All Sites** (N = 100,869) | **Site 1** (N = 8400) | **Site 2** (N = 45,165) | **Site 3** (N = 47,304) |
| Available at any time within 365 days before to 30 days after ibuprofen or indomethacin Initiation | | | | | | | | |
| Yes | 492,035 (53.2) | 13,781 (55.2) | 379,753 (59.5) | 98,501 (37.7) | 53,334 (52.9) | 4784 (57.0) | 31,449 (69.6) | 17,101 (36.2) |
| No | 432,621 (46.8) | 11,193 (44.8) | 258,976 (40.5) | 162,452 (62.3) | 47,535 (47.1) | 3616 (43.0) | 13,716 (30.4) | 30,203 (63.8) |
| Timing of availability | | | | | | | | |
| Before NSAID Initiation Only | 363,636 (39.3) | 10,393 (41.6) | 276,326 (43.3) | 76,917 (29.5) | 38,486 (38.2) | 3183 (37.9) | 22,437 (49.7) | 12,866 (27.2) |
| After NSAID Initiation Only | 120,992 (13.1) | 3518 (14.1) | 101,364 (15.9) | 16,110 (6.2) | 16,988 (16.8) | 1782 (21.2) | 10,926 (24.2) | 4280 (9.0) |
| Both before and after NSAID initiation | 58,793 (6.4) | 1980 (7.9) | 47,830 (7.5) | 8983 (3.4) | 8227 (8.2) | 696 (8.3) | 5131 (11.4) | 2400 (5.1) |
| | | | | | | | | |
| **Before Ibuprofen or Indomethacin initiation only [a, b, c]** | **All Sites (N = 363,636)** | **Site 1 (N = 10393)** | **Site 2 (N = 276,326)** | **Site 2 (N = 76,917)** | **All Sites (N = 38,486)** | **Site 1 (N = 3183)** | **Site 2 (N = 22,437)** | **Site 3 (N = 12,866)** |
| Emergency department | 55,278 (15.2) | 0 (0.0) | 55,278 (20.0) | 0 (0.0) | 5086 (13.2) | 0 (0.0) | 5086 (22.7) | 0 (0.0) |
| Inpatient | 23,639 (6.5) | 1381 (13.3) | 21,662 (7.8) | 596 (0.8) | 2993 (7.8) | 414 (13.0) | 2447 (10.9) | 132 (1.0) |
| Outpatient | 336,146 (92.4) | 9964 (95.9) | 249,583 (90.3) | 76,599 (99.6) | 36,627 (95.2) | 3044 (95.6) | 20,793 (92.7) | 12,790 (99.4) |
| | | | | | | | | |
| **After Ibuprofen or Indomethacin initiation only [a, b, c]** | **All Sites (N = 120,992)** | **Site 1 (N = 3518)** | **Site 2 (N = 101,364)** | **Site 3 (N = 16,110)** | **All Sites (N = 16,988)** | **Site 1 (N = 1782)** | **Site 2 (N = 10,926)** | **Site 3 (N = 4280)** |
| Emergency department | 27,290 (22.6) | 0 (0.0) | 27,290 (26.9) | 0 (0.0) | 1450 (8.5) | 0 (0.0) | 1450 (13.3) | 0 (0.0) |
| Inpatient | 8034 (6.6) | 760 (21.6) | 7100 (7.0) | 174 (1.1) | 695 (4.1) | 101 (5.7) | 544 (5.0) | 50 (1.2) |
| Outpatient | 95,458 (78.9) | 2963 (84.2) | 76,533 (75.5) | 15,962 (99.1) | 15,699 (92.4) | 1729 (97.0) | 9732 (89.1) | 4238 (99.0) |
| | | | | | | | | |
| **Both before and Ibuprofen or Indomethacin initiation [a, b, c]** | **All Sites (N = 58,793)** | **Site 1 (N = 1980)** | **Site 2 (N = 47,830)** | **Site 3 (N = 8983)** | **All Sites (N = 8227)** | **Site 1 (N = 696)** | **Site 2 (N = 5131)** | **Site 3 (N = 2400)** |
| Emergency and Emergency | 5403 (9.2) | 0 (0.0) | 5403 (11.3) | 0 (0.0) | 445 (5.4) | 0 (0.0) | 445 (8.7) | 0 (0.0) |
| Emergency and Inpatient | 2288 (3.9) | 0 (0.0) | 2288 (4.8) | 0 (0.0) | 275 (3.3) | 0 (0.0) | 275 (5.4) | 0 (0.0) |
| Emergency and Outpatient | 8026 (13.7) | 0 (0.0) | 8026 (16.8) | 0 (0.0) | 957 (11.6) | 0 (0.0) | 957 (18.7) | 0 (0.0) |
| Inpatient and Emergency | 1798 (3.1) | 0 (0.0) | 1798 (3.8) | 0 (0.0) | 218 (2.6) | 0 (0.0) | 218 (4.2) | 0 (0.0) |
| Inpatient and Inpatient | 2369 (4.0) | 198 (10.0) | 2104 (4.4) | 67 (0.7) | 298 (3.6) | 32 (4.6) | 252 (4.9) | 14 (0.6) |
| Inpatient and Outpatient | 3984 (6.8) | 249 (12.6) | 3686 (7.7) | 49 (0.5) | 561 (6.8) | 112 (16.1) | 429 (8.4) | 20 (0.8) |
| Outpatient and Emergency | 12064 (20.5) | 0 (0.0) | 12064 (25.2) | 0 (0.0) | 797 (9.7) | 0 (0.0) | 797 (15.5) | 0 (0.0) |
| Outpatient and Inpatient | 5420 (9.2) | 510 (25.8) | 4835 (10.1) | 75 (0.8) | 462 (5.6) | 51 (7.3) | 390 (7.6) | 21 (0.9) |
| Outpatient and Outpatient | 41,600 (70.8) | 1516 (76.6) | 31,201 (65.2) | 8883 (98.9) | 7075 (86.0) | 639 (91.8) | 4065 (79.2) | 2371 (98.8) |

| Hemoglobin Result Value Availability Before and After Ibuprofen or Indomethacin Initiation [a, b, c, d] | Ibuprofen | | | | Indomethacin | | | |
|---|---|---|---|---|---|---|---|---|
| | All Sites (N = 924,656) | Site 1 (N = 24,974) | Site 2 (N = 638,729) | Site 3 (N = 260,953) | All Sites (N = 100,869) | Site 1 (N = 8400) | Site 2 (N = 45,165) | Site 3 (N = 47,304) |

[a] Only Site 2 has emergency department laboratory test results available in the MSDD
[b] Outpatient and unknown settings are considered together because Mini-Sentinel Data Partners have stated that laboratory results with the "setting" variable populated with "unknown" are primarily outpatient laboratory test results
[c] The N in each setting do not add to the N for the "Any" setting or "Overall" numbers because the same individual could have had HGB results from more than one setting (i.e., same individual could be counted in different settings)

***Upper gastrointestinal bleeding outcomes in the ibuprofen and indomethacin sub-cohort***

As shown in the top section of **Table 53**, inpatient UGI bleeding diagnoses, regardless of whether or not a drop in HGB was available from laboratory results, accounted for 15.0% of the bleeding outcomes within 30 days after ibuprofen initiation and 33.3% of the bleeding outcomes after indomethacin initiation (group 1). Available HGB results (HGB drop $\geq$ 3 g/dL) together with an UGI bleeding diagnosis assigned in a non-inpatient setting identified < 1% of the UGI bleeding cases (group 2). A HGB drop $\geq$ 3 g/dL and no coded bleeding diagnosis identified 54.0% (N = 1253) of the bleeding cases in individuals taking ibuprofen and 23.8% (N = 118) of the bleeding cases in individuals taking indomethacin (group 3). Non-inpatient diagnoses without available HGB results or with a HGB drop < 3 g/dL accounted for 30.3% of cases among individuals taking ibuprofen and 42.3% of cases among individuals taking indomethacin (group 4).

Given that we previously determined only 9.5% of group 3 patients across the entire NSAID cohort had any non-UGI bleeding diagnosis within 30 days (section VIIF2b), the Workgroup did not include group 3 cases in the total of UGI bleeding cases for this sub-cohort initiating ibuprofen or indomethacin (bottom section of **Table 53**). Similar to the entire cohort, it is again clear that availability of HGB result values alters the estimated numbers (and rates) of UGI bleeding outcomes after ibuprofen or indomethacin initiation more at sites that have a higher percentage of the population with laboratory results available than at sites where laboratory results are missing in a higher percentage of the population.

**Table 53. Upper Gastrointestinal Bleeding Outcomes within 30 Days after initiating Ibuprofen or Indomethacin using Varied Outcomes Definitions, Overall and by Site**

| Group | Bleeding Outcomes Criteria (mutually exclusive categories) | Data Partner Site | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | All Sites | | Site 1 | | Site 2 | | Site 3 | |
| | | Ibuprofen | Indomethacin | Ibuprofen | Indomethacin | Ibuprofen | Indomethacin | Ibuprofen | Indomethacin |
| 1 | Inpatient diagnoses (with or without an observed HGB drop ≥ 3 g/dL) | 348 (15.0) | 165 (33.3) | 7 (9.1) | 4 (14.8) | 170 (10.9) | 39 (22.4) | 171 (24.8) | 122 (41.4) |
| 2 | Non-inpatient diagnosis with drop in HGB ≥ 3 g/dL | 16 (0.7) | 3 (0.6) | 0 (0.0) | 1 (3.7) | 14 (0.9) | 2 (1.1) | 2 (0.3) | 0 (0.0) |
| 3 | Observed drop in HBG ≥ 3 g/dL (no coded UGI bleeding diagnosis) | 1253 (54.0) | 118 (23.8) | 55 (71.4) | 15 (55.6) | 1125 (72.4) | 83 (47.7) | 73 (10.6) | 20 (6.8) |
| 4 | Non-inpatient diagnosis without observed drop in HGB | 702 (30.3) | 210 (42.3) | 15 (19.5) | 7 (25.9) | 244 (15.7) | 50 (28.7) | 443 (64.3) | 153 (51.9) |
| 1 - 4 | Total bleeding outcomes | 2319 | 496 | 77 | 27 | 1553 | 174 | 689 | 295 |
| **Group** | **Bleeding Outcomes Criteria (mutually exclusive categories)** | **All Sites** | | **Site 1** | | **Site 2** | | **Site 3** | |
| | | **Ibuprofen** | **Indomethacin** | **Ibuprofen** | **Indomethacin** | **Ibuprofen** | **Indomethacin** | **Ibuprofen** | **Indomethacin** |
| 1 | Inpatient diagnoses (with or without an observed HGB drop ≥ 3 g/dL) | 348 (32.6) | 165 (43.7) | 7 (31.8) | 4 (33.3) | 170 (39.7) | 39 (42.9) | 171 (27.8) | 122 (44.4) |
| 2 | Non-inpatient diagnosis with drop in HGB ≥ 3 g/dL | 16 (1.5) | 3 (0.8) | 0 (0.0) | 1 (0.8) | 14 (3.3) | 2 (2.2) | 2 (0.3) | 0 (0.0) |
| 4 | Non-inpatient diagnosis without observed drop in HGB | 702 (65.9) | 210 (55.6) | 15 (68.2) | 7 (58.3) | 244 (57.0) | 50 (54.9) | 443 (71.9) | 153 (55.6) |
| 1, 2, 4 | Total UGI bleeding outcomes without Group 3 | 1066 | 378 | 22 | 12 | 428 | 91 | 616 | 275 |

### d. Summary of Outcomes Detection Test Case 2

In the MSDD, from these three participating sites, HGB results were available at any time within 365 days before to 30 days after NSAID initiation for 45.3% of the cohort. HGB results availability varied by data partner site. Only 6.8% of the cohort had a HGB result available both within 365 days before to 30 days after NSAID initiation. Locations where HGB results were obtained at the sites that contribute inpatient laboratory results (integrated healthcare delivery systems) and emergency department laboratory results highlight the importance of having laboratory result values from the inpatient and emergency department care settings when an acute outcome such as UGI bleeding is assessed.

Using available HGB result values indicating a HGB drop ≥ 3 g/dL together with an UGI bleeding diagnosis assigned in a non-inpatient setting identified few additional cases. Use of HGB result values *alone* to identify cases of bleeding cannot be used to reliably detect UGI bleeding outcomes.

Across the entire cohort of patients identified as having an UGI bleeding outcome within 30 days after NSAID initiation using a coded inpatient diagnosis, only 13.9% were confirmed as having a HGB decrease ≥ 3 g/dL using HGB result values available both before and after NSAID initiation. An additional 15.6% were confirmed as having a HGB decrease < 3 g/dL. However, a high percentage of those individuals had CPT codes available both before and after NSAID initiation, indicating that HGB result values were missing in the MSDD but the HGB laboratory tests had been performed.

### 3. Summary of the performance of laboratory test results in outcomes detection

In these outcomes detection test cases, we assessed whether including laboratory results increased the number of individuals identified as developing the outcome of interest. The test case addressing UGI bleeding required assessing change in laboratory result values between baseline and follow-up time periods, while the test case addressing incident diabetes used follow-up laboratory result values only. Variability in missingness was again observed across sites.

In the first test case, laboratory results did not identify diabetes outcomes earlier but did identify additional outcomes at all three participating data partner sites. The rate increased the least at the large national insurer site, the site with the lowest percentage of the cohort with laboratory results available.

In the second outcomes test case, use of laboratory test results alone (i.e., without a coded diagnosis) was helpful for detecting a few additional bleeding outcomes, but was not helpful for detecting UGI bleeding outcomes (i.e., the specific body site required for the test case), and in some cases, the values of the clinical laboratory results were consistent with other outcomes (e.g., anemia of chronic disease). Availability of laboratory result values again altered the estimated numbers of outcomes after drug initiation more at sites that had a high percentage of the cohort with laboratory results available than at sites that had a higher percentage of the cohort with missing laboratory results.

In conclusion, in one outcome test case, including laboratory results identified an important number of additional outcomes, while in the other test case few additional outcomes were identified using laboratory results data. ***Based on results of these two test cases, laboratory test results should be used judiciously to supplement outcomes detection. The decision to include laboratory results as criteria for outcomes detection must be made on a scenario-by-scenario basis after consideration of the scenario-specific parameters and degree of missingness of the laboratory results.***

### 4. Considerations around censoring patients from the cohort if no outcome result value is observed

Censoring in longitudinal studies aims to remove individuals from a cohort when they are no longer at risk of the outcome or when the outcome cannot be expected to be captured. For a morbidity outcome, censoring at time of death or health plan disenrollment are typically straightforward decisions. In contrast, decisions to censor based on elapsed time without a laboratory result or without a medical visit are less clear (i.e., the laboratory result or the medical visit cannot be considered as an indicator in determining whether the individual has the outcome because lack of the outcome detection may reflect lack of care seeking).

Consider the scenario where a decision is made to censor patients 12 months after the medical product exposure if no laboratory results are observed within that time interval, but then the patient has the outcome observed at month 13. Should the patient have been censored at month 12? In general, it is

preferred to determine follow-up duration *a priori* based on biological plausibility rather than making a data driven decision later in the project.

Considerations around the decision to censor are not always straightforward and can vary depending on the nature of the outcome studied. Two idealized (extreme) examples help illustrate these points. In these examples, it is assumed that (similar to Data Partner Sites 1 and 2 in this methods work) laboratory results are reliably available in the MSDD if the laboratory test was completed.

1. Condition presents with symptoms. Laboratory test results are expected to then be gathered as part of the clinical assessment and therefore available in the dataset. In this condition, time without laboratory test results is meaningful – because it implies that the condition did not occur -- and the laboratory test results should be considered.
2. Condition is asymptomatic. The condition is only detected if laboratory tests are done, and laboratory tests are typically ordered as screening or monitoring tests rather than being prompted by clinical events or symptoms. In this condition, time without laboratory testing contributes little. It is reasonable that, if the individual does not have laboratory tests completed, the individual does not contribute any information and his or her follow up time should not contribute to the study. In this example, the decision of *when* to censor adds complexity because the laboratory test results may or may not provide information about *when* the condition occurred. The condition onset could have been at any time after the last "negative" laboratory test result or index date. Time to event analysis without considering this interval censoring is inappropriate. [99]

The idealized situations presented above often do not reflect real-life situations. There are some common complicating factors:

1. Few diseases fit the extremes of the idealized situations. Diabetes, for example, usually presents asymptomatically but occasionally presents with symptoms. However, many conditions present "close enough" to one of the idealized extremes that it can still be appropriate to model the condition in that way. For example, with diabetes it is not necessarily incorrect to model it as if it were always asymptomatic, particularly in the adult population.
2. Laboratory test results are often not used in isolation. That is, information about diagnosis codes and prescription medications are generally available in the observational data. How this information should be considered in the question of when to censor warrants future study.
3. What should be done when we know we do not have all laboratory results that were actually performed (e.g., differences in the percentages of patients with laboratory procedure claims codes and result values from the large national insurer Data Partners)? Should those evaluations focus on procedure claims codes and visits rather than on laboratory test results? How this information should be considered in the question of when to censor also warrants future study.

Determining time to censoring should take into account these considerations:

1. Relative timeframe required to develop the outcome (i.e., outcomes with short lag time after medical product exposure favor censoring if no observed laboratory result within the anticipated outcome timeframe),
2. Whether the outcome of interest is (usually) symptomatic or asymptomatic (i.e., symptomatic outcomes sometimes favor longer follow-up).

3. Length of follow-up (i.e., longer follow-up favors censoring if no laboratory testing or medical visit)
4. Healthcare utilization (i.e., percentage with and without and frequency of medical care during follow-up as a proxy for "opportunity" for laboratory testing and visits),
5. Completeness of capture of laboratory tests that were performed, (in general, making censoring and other decisions based on laboratory test result availability may be less advisable if capture of performed tests is poor).

We recommend considering censoring in settings when cohorts have relatively long follow-up (relative to the expected onset of an asymptomatic outcome of interest or relative to the expected onset of an acutely symptomatic outcome), and in cohorts where healthcare utilization is confirmed for the majority of the cohort within the timeframe of interest. A censored cohort could be used to conduct a sensitivity analysis (or sensitivity analyses of different time points) to ensure study conclusions are robust to censoring decisions. The censoring criteria could be similar to the following: censor at X months for individuals who, at that time, have not yet had either the outcome of interest or the laboratory test result of interest. An alternative criterion, especially relevant for symptomatic outcomes, could be to censor if the individual has not had a medical visit within X months (i.e., if an individual with the outcome of interest could reasonability be expected to have had the laboratory test completed if seen by a clinician at a medical visit).

## 5. Considerations around imputing outcomes

In general, it can be helpful to impute covariates. However, in general, it is not helpful to impute outcomes (and imputing outcomes is not often done). Consider the following dataset: The outcome is completely observed for all individuals in the cohort, but some covariate data are missing for some individuals. In this scenario, imputation of the covariates could increase efficiency because if an individual has an observed outcome and has some observed and some missing covariates data, there is information in that individual's data about the relationship(s) between the covariates and the outcome -- even for individuals who have just partial covariate data. For example, if there were ten covariates and an individual was missing data for one covariate, there would be information in that individual's data about the relationship between the outcome and the other nine covariates. In addition, individuals with complete data would provide information about the relationship between the missing covariate and the nine other covariates. If covariates adequately adjust for differences in who is missing the outcome, data are essentially MAR and MI, which assumes MAR data, can be used to impute the missing covariates. Now consider a dataset where the outcome is missing for some individuals in the cohort, but the covariates are completely observed. For an individual whose outcome is missing, his or her data contain information about the relationship between the covariates but no information about the relationship between the outcome and any of the covariates -- because the outcome is missing. All information about the relationship between the outcome and covariates must come from other individuals in the cohort. No efficiency is gained. In addition, analytic methods such as mixed models can be used to retain individuals with missing outcomes and account for missing outcome data if it is MAR.

Imputation of laboratory results to identify individuals with the outcome of interest is problematic from a clinical perspective. The frequency of laboratory test ordering can be influenced by the health status of the individual in general and the clinician's suspicion for the outcome of interest in particular. These are difficult factors to objectively measure. Laboratory test result values for the outcome of interest are likely to be MNAR in this scenario. With MNAR data, assumptions have to be made about parameters

that are not identifiable from the observed data. For example, we might make assumptions about how individuals with missing laboratory results data differ from individuals with complete data; these individuals may differ in ways that are not explained by their observed data. The key point is that assumptions must be made in absence of observed data.

Most MNAR methods are relatively complex, but these methods are becoming more accessible. Some MNAR methods look within subgroups that have different sources of missing data. In the Mini-Sentinel setting, data partner sites are subgroups that can have different sources of missing data. While exploration of MNAR methods was beyond the scope of this project, the baseline confounder test case completed as part of this project that considered missing data models overall and then by sites gives a rough example of dealing with subgroup missing data models versus combined models.

With data collected as part of routine healthcare delivery, one cannot be sure missing data are MCAR, MAR or MNAR.[100] In all likelihood, with data such as those contained in the MSDD, and based on the findings from work conducted by this Workgroup, missing laboratory results data are never completely MCAR, MAR, or MNAR, but rather are a mixture of missing observations that are completely random and missing observations that depend on either observed data or unobserved data.

# VIII. RECOMMENDATIONS AND SPECIFICATIONS FOR INCORPORATING MINI-SENTINEL DISTRIBUTED DATABASE LABORATORY RESULTS DATA INTO MINI-SENTINEL SAFETY ANALYSES

The Workgroup selected and tested methods for use when analyzing observational healthcare administrative, EHR, and other clinical databases, with specific attention to clinical laboratory results when missing data are expected. In the context of medical product safety surveillance in the MSDD environment, the Workgroup evaluated the use of laboratory results for baseline confounding adjustment, cohort identification, and detecting health outcomes. This section of the report provides recommendations and specifications for incorporating laboratory results data into Mini-Sentinel safety analyses. Based on findings from Workgroup activities, we discuss whether we can increase use of these data for various purposes, provide recommendations on strategies for increasing use of laboratory results data, and discuss whether and how laboratory results data can be incorporated into the Mini-Sentinel Routine Analytic Framework and Protocol-Based Assessments (PBA).

## A. AVAILABILITY OF LABORATORY RESULTS DATA IN MSDD

In section III A we provided an overview of the availability of laboratory results data in MSDD and sources of missing data. In consultation with FDA investigators in the Workgroup, seven missing data test cases (3 baseline confounder, 2 cohort identification, and 2 health outcomes) were selected (section VII B). Each test case was essentially a separate cohort study. Three representative Data Partners provided data for the test cases (section VII C).

In considering use of laboratory results data for medical product safety surveillance, important considerations are recognizing that missingness exists and that the level of missing data differs substantially across Data Partners. To appropriately use laboratory results data, steps must be taken to assess the extent of missing data for the specific scenario, to characterize the missingness by site, and to apply statistical tools that aid in managing and minimizing the missing data challenges (discussed further below).

## B. USE OF LABORATORY RESULTS DATA FOR BASELINE CONFOUNDER ADJUSTMENT

Based on the findings from the baseline confounder test cases (section VII D), we list below how laboratory results data can be used for baseline confounder adjustment to support medical product safety surveillance.

### 1. Improvements in confounding adjustment

Findings from these test cases indicated that laboratory test results were strongly associated with outcomes but had limited impact as confounder adjustments. There was great variability in missing data patterns for specific laboratory test results and across sites. Therefore, we do not generalize our finding that laboratory test result adjustments will not have substantial impacts on other cohorts or other analyses. Additional scenarios could be explored in which laboratory test results are, for example, strong baseline confounders.

### 2. Performance of missing data methods

Different missing data methods often provided comparable "answers." Comparable results were particularly evident in comparisons of multiple imputation (MI) methods. Missing data estimated by MI using either predictive mean matching (PMM) or regression methods resulted in identical or very similar point and CI estimates (i.e., differing by no more than 0.01 in both test cases 1 and 2). PMM produced wider CIs. Test case 2 results also suggested that the performance of missing data methods (MI regression method, MI PMM, vs. indicator variable method) was similar. Indicator variable method and PMM produced similar results in test case 3. However, models that ignored missing data (especially in site-specific analyses) or complete case models gave results different from models using MI methods. This was somewhat expected because complete case models assume data are missing completely at random (MCAR), whereas laboratory test results data can be MCAR, missing at random (MAR), missing not at random (MNAR), or a combination of MCAR, MAR, and MNAR; therefore employing complete case analysis with MAR or MNAR data provides biased estimates.

Differences in run times are important considerations. We employed SAS® version 9.3 or 9.4 for all analyses. The MI PMM and the MI regression methods had large differences in run times. As an example, both these methods were used to impute missing creatinine values in baseline confounder adjustment test case 2. This cohort had 198,265 members with 45% missing creatinine. MI using the regression method in SAS® version 9.4 required less than five minutes of run time and less than one minute of CPU time. By contrast, MI PMM was substantially less efficient; run time was over five hours and required 23 minutes of CPU time. Given the comparable performance in results and differences in computational efficiency, we recommend using the MI regression method. We do not know whether such dramatic differences in run time would occur with other software packages.

### 3. Site differences

Covariate associations with missing data varied across sites. Missing data methods need to allow for this variability, for example, by conducting imputations within sites or by employing models that include site by variable interaction terms.

Given differences across sites in missingness, cohort characteristics, and predictors of missingness, site-specific estimates and analytic methods that allow for site differences in associations with treatment and outcomes are recommended over a single pooled analysis with site indicators for adjustment. For

example, wider CIs were observed in meta-analyses that combined site-specific imputed results compared to pooled data analyses (test case 1), likely influenced by varied associations in the site-specific imputations models.

### 4. Other considerations

The literature review by the Workgroup (section IV) offers important lessons:

- Most literature relating to laboratory results data addresses Patient-Level or Provider-Level missing data. Little is known about how to handle Facility-Level, Organization-Level, Care-Setting Level, and Temporal-Level Missingness.
- Methods with known limitations, such as simple mean or median replacement of missing values that result in inappropriately small standard errors, should not be considered for Mini-Sentinel activities.
- Consistent patterns of missing laboratory results data should not be expected across studies. Missing data patterns are influenced by the specific laboratory test(s) of interest, study population characteristics, and provider and system factors. Describing missing data overall and by key variables within the study population is an important early step in studies.

The complexities of reasons for missing laboratory results data deserve thoughtful discussion during analysis planning. Any missing data technique will produce biased estimates if the underlying assumptions about the data are violated. MCAR, MAR, and MNAR are assumptions that govern the performance of missing data handling strategies.[2] Complete case analysis assumes and requires MCAR data, which rarely holds for missing laboratory results and should not be used. MAR patterns may be found, but are difficult to confirm. MI, a common missing data technique, requires the less stringent MAR assumption. This is one reason MI techniques are recommended. Whether there is evidence of a strong MNAR pattern should be considered.

## C. USE OF LABORATORY RESULTS DATA FOR COHORT IDENTIFICATION

In the cohort identification test cases, we assessed whether using laboratory results enhanced identification of a cohort, either increasing the size of the cohort or identifying individuals for cohort inclusion at an earlier date than they would have been identified using only diagnoses and procedures. Both cohort identification test cases showed that cohort sizes were augmented through inclusion of laboratory results; one test case also suggested earlier identification of individuals occurs when cohort inclusion uses laboratory results data in addition to diagnosis and procedure codes. However, given the variability in missing data patterns for specific laboratory results and across sites, it is again possible that these results may not generalize to other cohorts or other analyses. The magnitude of gain in cohort size depends on the specific laboratory test and the specific cohort. For routine Mini-Sentinel work, we recommend considering the use of laboratory results data to assist in cohort identification in combination with algorithms based on diagnosis and procedure codes and medications. Studies should adjust for site variations and further study the subgroup of individuals in the cohort who only have a laboratory-based indicator in the observational database, in particular within the context of implications related to risks of drug exposures.

Characterizing the population helps inform misclassification bias. This is because laboratory data are ‘opportunistic,’ that is, data are available (or missing) for reasons that may not be known by, and beyond the control of, the investigator. Characterizing individuals who are only included in the cohort by

use of laboratory results can clarify whether those individuals should be included. At a minimum, sensitivity analyses with and without such individuals are recommended.

## D. USE OF LABORATORY RESULTS DATA FOR OUTCOMES DETECTION

In the outcomes detection test cases we assessed whether including laboratory test results increased the number of individuals identified as developing the outcome of interest. These test cases required assessing change in laboratory result values between baseline and follow-up time periods. Again, given the variability in missingness for specific laboratory results and across sites, our findings may or may not generalize to other cohorts or analyses.

In the first outcomes test case, including laboratory result values did not identify diabetes outcomes earlier but did identify additional outcomes. At all three sites, the outcome rates increased when diagnosis codes, medications, and laboratory result values were considered versus considering only diagnosis codes and medications. The rate increased the least at the large national insurer, the site with the lowest proportion of available laboratory result values.

In the second outcomes test case, availability of laboratory result values again altered the numbers of outcomes after drug initiation more at sites that had a high proportion of the cohort with laboratory results available than at sites with a higher proportion of missing laboratory results. However, in the second outcomes test case, use of laboratory results alone was not useful for detecting the bleeding outcome at the body site specific to the test case (upper gastrointestinal tract).

The differences between the usefulness of the laboratory results in the two outcomes test cases is related to the "specificity" of the laboratory test types. In the first outcomes test case, glucose was the laboratory test of interest. Glucose laboratory result values are part of the diagnostic pathway of the diabetes outcome, an outcome that is not body site specific. In the second outcomes test case, hemoglobin was the laboratory test of interest. While hemoglobin result values are part of the diagnostic pathway of the bleeding outcome, the body site where the bleeding occurred is also an important component of this outcome. In the first test case, use of the glucose laboratory results alone, when the result value was above a known threshold, was sufficient to indicate the outcome. In the second test case, use of change in the hemoglobin laboratory results alone was insufficient to indicate the outcome because evidence of the specific body site of bleeding was also required.

Defining outcomes using diagnosis codes, procedures, and laboratory result values can introduce inconsistency across sites if codes, etc., are available or missing for reasons unknown to the investigator. Missingness of laboratory results can further accentuate site inconsistencies because of differing practices and laboratory results data availability. Thus, using laboratory results can contribute false negatives and false positives, although using laboratory results potentially can contribute more false negative and false positives than diagnosis codes.

From a statistical perspective, it is often appropriate and helpful to impute covariates. However, it usually is neither appropriate nor helpful to impute outcomes and, therefore, we do not recommend imputing them (section VII F 5).

We make no general recommendation about employing laboratory results data to detect health outcomes. Decisions about whether or not to include laboratory results in algorithms that identify health outcomes in the MSDD should be made on a scenario-by-scenario basis.

## E. RECOMMENDATION FOR THE GENERAL PROCESS FOR INCORPORATING LABORATORY RESULTS DATA INTO MINI-SENTINEL SAFETY ANALYSES

We recommend the following approach to include laboratory results data in Mini-Sentinel safety analyses. Aspects of this approach may include automated analyses that are already part of Mini-Sentinel's Routine Analytic Framework; other aspects require enhancements to existing tools for conducting routine analyses (discussed in the next section).

### 1. Baseline confounder adjustment

- Describe missing data in the study cohort is an important step in considering missing data technique options overall and by site. Characterizing the population helps inform potential bias. Characterization of missing laboratory results data can clarify why data are available for certain people and not others.
- Investigate predictors of missing laboratory results values overall and by site.
- Select approaches to handling missing laboratory results data. Three approaches were compared by this Workgroup: MI by regression, MI by PMM, and MI by indicator variable method. The MI regression method is common and easy to implement. Because MI approaches rely on a normality assumption, skewness should be assessed up front and a log-transformation should be performed to improve symmetry if needed.
- Compare results between different strategies for handling laboratory results data (e.g., analyses that do not include laboratory results at all in the analysis and analyses that account for missing laboratory results using missing data techniques such as MI). The primary comparisons are the estimated coefficients and 95% CI for the exposures of interest. The first question is whether including laboratory results makes any difference. Consistent results enhance confidence in the findings, while differing results support the importance of using missing data techniques. Regardless, examination of varied models will enhance understanding of the importance of the laboratory results data.

### 2. Cohort identification

- Characterizing the population helps inform bias overall and by site (and across years if needed). Contrasting those identified by diagnosis or procedure codes alone, by laboratory results alone, or by either diagnosis or procedures codes or laboratory results should be performed during an early phase of the project.
- In certain assessments, it can be useful to assess whether laboratory results enable cohort identification and inclusion earlier, apart from gains in cohort size, and to examine agreement between different cohort identification algorithms. These activities inform the tradeoffs between enhanced cohort identification and misclassification.

### 3. Outcome detection

- Characterize the population by exposure and by site.
- Assess the availability of baseline laboratory result values across the cohort (by type of test as appropriate).

- Assess the availability of follow-up laboratory result values across the cohort (by type of test as appropriate).
- Examine numbers of outcomes detected, contrasting different outcomes definitions (e.g., diagnosis, medication, laboratory results).
- Compare the time to outcome detection, contrasting when laboratory results were and were not used for outcome detection.
- Examine outcome rates for each site, stratified by whether or not any follow-up laboratory result value was available. Outcome rates may be higher among individuals with follow-up laboratory results available (regardless of whether the outcome was determined using laboratory results values), suggesting bias in selecting which individuals received laboratory testing.
- Assess relationships between baseline characteristics and missing follow-up laboratory result values.
- Assess the effect of the various outcome definitions on relationships between individual drug exposures and the first indicator of the outcome overall and by site

## F. RECOMMENDATIONS FOR INCORPORATING LABORATORY RESULTS DATA INTO MINI-SENTINEL'S ROUTINE ANALYTIC FRAMEWORK

The Mini-Sentinel Cohort Identification and Descriptive Analysis (CIDA) tool is the foundation of the Mini-Sentinel Routine Analytic Framework. CIDA supports identifying, extracting, and characterizing cohorts from the MSDD based on specifying requestor-defined options. This Workgroup employed the CIDA tool for test case cohort development when feasible. CIDA was not yet finalized when the data were extracted for the first baseline confounder adjustment test case, and the first outcome test case used the same dataset as the first baseline confounder adjustment test case. CIDA was not used in those two test cases. Across the other five test cases, CIDA was used to identify four of the five cohorts.

A challenge that precluded successful use of the Mini-Sentinel Routine Analytic Framework by this Workgroup was found when identifying the cohort for the baseline confounder test case that examined antimicrobial use among warfarin users. Cohort inclusion required an index warfarin date as well as an index antimicrobial date and also required temporal sequencing of those dates. The existing CIDA tool cannot handle multiple index dates or temporal sequencing. Custom coding was required to address those issues. Also, in that test case, covariate data returned using the CIDA tool were incorrectly anchored on the warfarin index date rather than (correctly anchored) on the antimicrobial index date. This limitation could not be remedied using existing Routine Analytic Framework capabilities. Based on what we learned from the warfarin baseline confounder test case, the Workgroup did not use the Mini-Sentinel Routine Analytic Framework to identify the cohort for the CKD cohort identification test case because that test case also required more than one index date. For scenarios with complex index dates, custom coding to identify the cohort of interest is currently required.

Custom coding was also required to explore the different methods of handling missing laboratory results data and to explore the different analytic techniques used in the test cases. Custom coding again was used in part because the CIDA tool first became available after commencement of this project; this project was one of the first to apply the CIDA tool. CIDA now has a module available for one of the analytic techniques used by this Workgroup (PS computation and matching).

The existing Mini-Sentinel Routine Analytic Framework has no modules available to date that describe missing data overall or by site or for applying missing data methods. Therefore, Mini-Sentinel Routine Analytic Framework modules would need to be developed to incorporate missing data techniques. Logical first steps would be to create a repeatable, reusable process to describe missing data within a cohort overall and by site to allow investigators to examine whether available laboratory test results data can be assumed to be MAR. This process would inform whether it is reasonable to impute the missing data. We recommend these steps, perhaps through forming a small workgroup charged with writing detailed functional specifications and processes to use CIDA to examine laboratory test results data when clinical laboratory data elements are required for a query.

Other potential enhancements using SAS®, include developing a module that could use PROC MI to impute missing data and PROC MIANALYZE to combine model results from the multiple imputed datasets. Different techniques in PROC MI would be needed if multiple variables had missing data (versus only the one laboratory test result type we studied in each test cases examined by the Workgroup). In addition, methods to combine the model results will differ for varied analytic models (i.e., data to be saved and pulled into PROC MIANALYZE differs in the items and naming). Further, it would be preferable if investigators had an opportunity to examine predictors of the missing data and postulate whether MAR is a reasonable possibility. Despite these complexities, a reasonable step, if prioritized as an enhancement would be to develop a module that would allow missing data to be imputed and combined using these SAS® procedures for analytic techniques commonly in use. This module could be used as a sensitivity analysis to examine if conclusions are impacted.[100, 101]

If multi-site laboratory results data enhanced by MI are to be routinely used in the Mini-Sentinel Routine Analytic Framework, enhancements to existing modules would be necessary. For example, the existing IPTW module in the Mini-Sentinel Routine Analytic Framework works with binary data, using Maentel–Haenzel techniques to perform meta-analysis on site-specific PS-based outcome regression models. These enhanced modules would need to be aligned to enable use of output from the imputation steps (i.e., final summarized data from PROC MIANALYZE).

Existing Mini-Sentinel query tools cannot utilize laboratory result values over time or changes in laboratory result values over time (i.e., sequential laboratory test result values). Assessing increases or decreases in laboratory result values is important for medical product safety surveillance, particularly for outcomes detection (assessed by this Workgroup) and for time-varying confounding (not in the scope of work of this Workgroup). If prioritized by the FDA, enhancing the existing Mini-Sentinel Routine Analytic Framework to enable assessment of multiple laboratory values over time and changes in the laboratory result values over time should be considered. The custom programming code used by this Workgroup could again inform the development of programming specifications to assess sequential laboratory result values, if this capability was prioritized for development.

## G. RECOMMENDATIONS FOR INCORPORATING LABORATORY RESULTS DATA INTO MINI-SENTINEL'S PROTOCOL-BASED ASSESSMENTS

Given existing capabilities of the Mini-Sentinel Routine Analytic Framework, use of laboratory results data in current Mini-Sentinel evaluations will often be feasible only within PBA. As the capabilities of the Mini-Sentinel Routine Analytic Framework further develop, laboratory results data may be more feasible to include in routine automated queries.

Each of the seven test case scenarios examined by this Workgroup was essentially a separate retrospective observational cohort study, whose purposes were methods development and examination. When viewed in that context, one can think of these seven test cases as seven separate PBAs.

All customized programming code used for these test cases was prepared, tested, and quality-checked in accordance with Mini-Sentinel policies and procedures. Therefore, code developed by this Workgroup could be shared with other workgroups conducting PBAs. However, because the custom coding used by this Workgroup only underwent a portion of the quality checking process to which code developed for a PBA is usually subjected (because the custom coding in this project was for methods development, not a true PBA), the existing custom coding likely would need additional testing to be in compliance with Mini-Sentinel Standard Operating Procedures depending on how and where the customized code was to be used in a future PBA.

Additionally, the custom code developed by the Workgroup could be modularized for use with other types of laboratory test results, in conjunction with other exposures or other outcomes of interest. For instance, the programming code sets the Workgroup developed that identified complex index dates and temporal sequencing could be modularized for future use. Similarly, the programming code the workgroup developed that assessed change in laboratory result values over time could also be modularized for future use.

## H. CONCLUSIONS ABOUT INCORPORATING MINI-SENTINEL DISTRIBUTED DATABASE LABORATORY RESULTS DATA INTO MINI-SENTINEL SAFETY ANALYSES

In conclusion, in all test cases, there were substantial differences in missing data across different Data Partners. This must be taken into consideration when using laboratory test results data.

Laboratory test results were strongly associated with outcomes but had limited impact as baseline confounder adjustments. Given the variability in missing data patterns for specific laboratory test results and across sites, we cannot ensure the generalizability of our finding that laboratory result adjustments do not have substantial impacts when assessed in other cohorts or in other analyses. If a priority for the FDA, exploring additional scenarios using laboratory results as baseline confounder adjustment and exploring time-varying confounders should be considered. For time-varying confounders, enhancing the existing Mini-Sentinel Routine Analytic Framework to enable use of multiple result values of the same laboratory test result over time would be necessary.

For routine Mini-Sentinel queries, where applicable, we recommend use of laboratory results data for cohort identification, supplementing algorithms based on diagnosis and procedure codes and medications. It will be important to adjust for site variations and to examine the subgroup of individuals who only have a laboratory-based indicator in the database, in particular within the context of bias and implications related to risks of drug exposures. The existing Mini-Sentinel Routine Analytic Framework

can be used for cohort identification, with the exception of situations in which multiple exposures and temporal relationships are required.

Not surprisingly, the availability of laboratory results data identified additional health outcomes of interest. If it is a priority for the FDA to use laboratory test results to assess outcomes, the existing Mini-Sentinel Routine Analytic Framework would require enhancement to enable assessing change in laboratory result values over time. Outcome detection using laboratory results values in addition to diagnosis and procedure codes (and medications) must also consider the substantial variation in missing data across sites.

In summary, the existing Mini-Sentinel Routine Analytic Framework can enable identification some cohorts that utilize laboratory results data. However, the Mini-Sentinel Routine Analytic Framework cannot assist with missing data characterization. Other tasks required to use laboratory results currently require custom programming. Therefore, the overall conclusion of the Workgroup is that further enhancement of existing Mini-Sentinel Routine Analytic Framework is required prior to routine use of laboratory results data in query fulfillment. At the current time, use of laboratory results data can be primarily accomplished in Protocol-Based Assessments.

# IX. REFERENCES


1. Raebel MA, Haynes K, Woodworth TS et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf* 2014;23(6):609-618.
2. Enders CK. *Applied missing data analysis*. New York: The Guilford Press; 2010.
3. Curtis LH, Weiner MG, Boudreau DM et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 1:23-31. doi: 10.1002/pds.2336.
4. Maro JC, Platt R, Holmes JH et al. Design of a national distributed health data network. *Ann Intern Med* 2009;151(5):341-344.
5. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010;48(6 Suppl):S45-S51. doi: 10.1097/MLR.0b013e3181d9919f.
6. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. *Clin Pharmacol Ther* 2011;90(6):883-887. doi: 10.1038/clpt.2011.236.
7. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. *N Engl J Med* 2011;364(6):498-499. doi: 10.1056/NEJMp1014427.
8. Platt R, Carnahan RM, Brown JS et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 1:1-8. doi: 10.1002/pds.2343.
9. Raebel MA, Chester EA, Newsom EE et al. Randomized trial to improve laboratory safety monitoring of drug therapy in ambulatory patients. *Pharmacother* 2006;26:619-626.
10. Raebel MA, Lyons EE, Bodily MA et al. Improving laboratory monitoring at initiation of drug therapy: A randomized trial. *Arch Intern Med* 2005;165:2395-2401.
11. Olson KL, Rasmussen J, Sandhoff BG, Merenich JA, for the Clinical Pharmacy Cardiac Risk Service (CPCRS) Study Group. Lipid management in patients with coronary artery disease by a clinical pharmacy service in a group model health maintenance organization. *Arch Intern Med* 2005;165:49-54.
12. McGinnis BD, Olson KL, Delate TMA, Stolcpart RS. Statin adherence and mortality in patients enrolled in a secondary prevention program. *Am J Manag Care* 2009;15(10):689-695.
13. Jaffe MG, Lee GA, Young JD, Sidney S, Go AS. Improved blood pressure control associated with a large-scale hypertension program. *JAMA* 2013;310(7):699-705.
14. Witt DM, Sadler MA, Shanahan RL, Mazzoli G, Tillman DJ. Effect of a centralized clinical pharmacy anticoagulation service on the outcomes of anticoagulation therapy. *Chest* 2005;127(5):1515-1522.
15. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581-592.
16. Filion KB, Joseph L, Boivin JF, Suissa S, Brophy JM. Trends in the prescription of anti-diabetic medications in the United Kingdom: a population-based analysis. *Pharmacoepidemiology & Drug Safety* 2009;18(10):973-976.
17. Morrato EH, Druss BG, Hartung DM et al. Small area variation and geographic and patient-specific determinants of metabolic testing in antipsychotic users. *Pharmacoepidemiology & Drug Safety* 2011;20(1):66-75.

18. Li J, Motsko SP, Goehring EL, Jr., Vendiola R, Maneno M, Jones JK. Longitudinal study on pediatric dyslipidemia in population-based claims database.[Erratum appears in Pharmacoepidemiol Drug Saf. 2010 Feb;19(2):216]. *Pharmacoepidemiology & Drug Safety* 2010;19(1):90-98.
19. LaFrance JP, Miller DR. Selective and non-selective non-steroidal anti-inflammatory drugs and the risk of acute kidney injury. *Pharmacoepidemiol Drug Saf* 2009;18(10):923-931.
20. Hung AM, Roumie CL, Greevy RA et al. Kidney function decline in metformin versus sulfonylurea initiators: assessment of time-dependent contribution of weight, blood pressure, and glycemic control. *Pharmacoepidemiol Drug Saf* 2013;22(6):623-631.
21. van Bruggen R, Gorter K, Stolk RP, Zuithoff P, Klungel OH, Rutten GE. Refill adherence and polypharmacy among patients with type 2 diabetes in general practice. *Pharmacoepidemiol Drug Saf* 2009;18(11):983-991.
22. Neugebauer R, Fireman B, Roy JA, O'Connor PJ, Selby JV. Dynamic marginal structural modeling to evaluate the comparative effectiveness of more or less aggressive treatment intensification strategies in adults with type 2 diabetes.[Erratum appears in Pharmacoepidemiol Drug Saf. 2012 Dec;21(12):1361]. *Pharmacoepidemiology & Drug Safety* 2012;21:Suppl-113.
23. Davidson MH, Jones PH. Retrospective comparison of the effectiveness of a fenofibrate 145 mg formulation compared with the standard 160 mg tablet. *Clin Drug Investig* 2008;28(10):615-623.
24. Pandya BJ, Bron M, McCall T et al. Achieving glycemic goal with initial versus sequential combination therapy using metformin and pioglitazone in type 2 diabetes mellitus. *Curr Med Res Opin* 2011;27(1):189-195.
25. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. *J Biomed Inform* 2014.
26. Schneeweiss S, Rassen JA, Glynn RJ et al. Supplementing claims data with outpatient laboratory test results to improve confounding adjustment in effectiveness studies of lipid-lowering treatments. *BMC Med Res Methodol* 2012;12:180.
27. Siew ED, Peterson JF, Eden SK, Moons KG, Ikizler TA, Matheny ME. Use of multiple imputation method to improve estimation of missing baseline serum creatinine in acute kidney injury research. *Clinical Journal of The American Society of Nephrology: CJASN* 2013;8(1):10-18.
28. Johnson ES, Thorp ML, Yang X, Charansonney OL, Smith DH. Predicting renal replacement therapy and mortality in CKD. *Am J Kidney Dis* 2007;50(4):559-565.
29. Wagner M, Ansell D, Kent DM et al. Predicting mortality in incident dialysis patients: an analysis of the United Kingdom Renal Registry. *American Journal of Kidney Diseases* 2011;57(6):894-902.
30. Faries D, Peng X, Pawaskar M, Price K, Stamey JD, Seaman JW, Jr. Evaluating the impact of unmeasured confounding with internal validation data: an example cost evaluation in type 2 diabetes. *Value in Health* 2013;16(2):259-266.
31. Witt DM, Delate T, Clark NP et al. Nonadherence with INR monitoring and anticoagulant complications. *Thrombosis Research* 2013;132(2):e124-e130.
32. Li J, Motsko SP, Goehring EL, Jr., Tave A, Pezzullo JC, Jones JK. Prevalence of pediatric dyslipidemia: comparison of a population-based claims database to national surveys. *Pharmacoepidemiology & Drug Safety* 2010;19(10):1031-1040.
33. Trickey AW, Fox EE, del Junco DJ et al. The impact of missing trauma data on predicting massive transfusion. *Journal of Trauma and Acute Care Surgery* 2013;75(1:Suppl 1):Suppl-74.
34. Raebel MA, Penfold R, McMahon AW et al. Adherence to guidelines for glucose assessment in starting second-generation antipsychotics. *Pediatrics* 2014;Published online October 6, 2014; DOI: 10.1542/peds.2014-0828.

35. Reddy SR, Ross-Degnan D, Zaslavsky AM, Soumerai SB, Wharam JF. Impact of a high-deductible health plan on outpatient visits and associated diagnostic tests. *Med Care* 2014;52(1):86-92.
36. Tomasson G, Peloquin C, Mohammad A et al. Risk for cardiovascular disease early and late after a diagnosis of giant-cell arteritis: a cohort study. *Ann Intern Med* 2014;160(2):73-80.
37. Mulla ZD, Seo B, Kalamegham R, Nuwayhid BS. Multiple imputation for missing laboratory data: an example from infectious disease epidemiology. *Ann Epidemiol* 2009;19(12):908-914.
38. Bansi L, Gazzard B, Post F et al. Biomarkers to monitor safety in people on art and risk of mortality. *Journal of Acquired Immune Deficiency Syndromes: JAIDS* 2012;60(1):51-58.
39. Richardson P, Henderson L, Davila JA et al. Surveillance for hepatocellular carcinoma: development and validation of an algorithm to classify tests in administrative and laboratory data. *Digestive Diseases & Sciences* 2010;55(11):3241-3251.
40. Nakano C, Hamano T, Fujii N et al. Combined use of vitamin D status and FGF23 for risk stratification of renal outcome. *Clinical Journal of The American Society of Nephrology: CJASN* 2012;7(5):810-819.
41. Nakano C, Hamano T, Fujii N et al. Intact fibroblast growth factor 23 levels predict incident cardiovascular event before but not after the start of dialysis. *Bone* 2012;50(6):1266-1274.
42. Raebel MA, Xu S, Goodrich GK et al. Initial antihyperglycemic drug therapy among 241,327 adults with newly identified diabetes from 2005 through 2010: a SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM) study. *Ann Pharmacother* 2013;47(10):1280-1291.
43. Waljee AK, Mukherjee A, Singal AG et al. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* 2013;3(8).
44. Badawi O, Breslow MJ. Readmissions and death after ICU discharge: development and validation of two predictive models. *PLoS One* 2012;7(11):e48758.
45. Harris LT, Koepsell TD, Haneuse SJ, Martin DP, Ralston JD. Glycemic control associated with secure patient-provider messaging within a shared electronic medical record: a longitudinal analysis. *Diabetes Care* 2013;36(9):2726-2733.
46. Cerda J, Cerda M, Kilcullen P, Prendergast J. In severe acute kidney injury, a higher serum creatinine is paradoxically associated with better patient survival. *Nephrology Dialysis Transplantation* 2007;22(10):2781-2784.
47. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* 2003;56(1):28-37.
48. Bejot Y, Aboa-Eboule C, Hervieu M et al. The deleterious effect of admission hyperglycemia on survival and functional outcome in patients with intracerebral hemorrhage. *Stroke* 2012;43(1):243-245.
49. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 2014;14:75.
50. Lombard C, Deeks A, Jolley D, Ball K, Teede H. A low intensity, community based lifestyle programme to prevent weight gain in women with young children: cluster randomised controlled trial. *BMJ* 2010;341(c3215).
51. Holbrook A, Pullenayegum E, Thabane L et al. Shared electronic vascular risk decision support in primary care: computerization of medical practices for the enhancement of yherapeutic effectiveness (COMPETE III) randomized trial. *Arch Intern Med* 2011;171(19):1736-1744.
52. de Jong WJ, Stoepker L, Nellensteijn DR, Groen H, El MM, Hulscher JB. External validation of the blunt abdominal trauma in children (BATiC) score: Ruling out significant abdominal injury in children. *The Journal of Trauma and Acute Care Surgery* 2014;76(5):1282-1287.

53. Mammone A, Pezzotti P, Angeletti C et al. HIV incidence estimate combining HIV/AIDS surveillance, testing history information and HIV test to identify recent infections in Lazio, Italy. *BMC Infectious Diseases* 2012;12(65).
54. Serrat C, Gomen G, Garcia de OP, Cayla JA. CD4+ lymphocytes and tuberculin skin test as survival predictors in pulmonary tuberculosis HIV-infected patients. *Internation J Epidemiol* 1998;27(4):703-712.
55. van Mourik MS, Groenwold RH, Berkelbach van der Sprenkel JW, van Solinge WW, Troelstra A, Bonten MJ. Automated detection of external ventricular and lumbar drain-related meningitis using laboratory and microbiology results and medication data. *PLoS One* 2011;6(8):e22846.
56. Amarasingham R, Moore BJ, Tabak YP et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care* 2010;48(11):981-988.
57. Polsky D, Eremina D, Hess G et al. The importance of clinical variables in comparative analyses using propensity-score matching: the case of ESA costs for the treatment of chemotherapy-induced anaemia. *Pharmacoeconomics* 2009;27(9):755-765.
58. Parsons HM, Henderson WG, Ziegenfuss JY, Davern M, Al-Refaie WB. Missing data and interpretation of cancer surgery outcomes at the American College of Surgeons National Surgical Quality Improvement Program. *J Am Coll Surg* 2011;213(3):379-391.
59. Ho SC, Chen YM, Ho SS, Woo JL. Soy isoflavone supplementation and fasting serum glucose and lipid profile among postmenopausal Chinese women: a double-blind, randomized, placebo-controlled trial. *Menopause* 2007;14(5):905-912.
60. Shardell M, El-Kamary SS. Sensitivity analysis of informatively coarsened data using pattern mixture models. *H Biopharm Stat* 2009;19:1018.
61. Neugebauer R, Fireman B, Roy J, O'Connor PJ, Selby JV. Dynamic marginal structural modeling to evaluate the comparative effectiveness and safety of more or less aggressive treatment intensification strategies in type 2 diabetes. *Pharmacoepidemiol Drug Saf* 2012;21:99-113.
62. Finkelstein DM, Wang R, Ficociello LH, Schoenfeld DA. A score test for association of a longitudinal marker and an event with missing data. *Biometrics* 2010;66(3):726-732.
63. Schoenfeld DA, Rajicic N, Ficociello LH, Finkelstein DM. A test for the relationship between a time-varying marker and both recovery and progression with missing data. *Stat Med* 2011;30(7):718-724.
64. De Hert M, Detraux J, van Winkel R, Yu W, Correll CU. Metabolic and cardiovascular adverse effects associated with antipsychotic drugs. *Nat Rev Endocrinol* 2011;8(2):114-126.
65. Newcomer JW. Second-generation (atypical) antipsychotics and metabolic effects: a comprehensive literature review. *CNS Drugs* 2005;19 Suppl 1:1-93.
66. Correll CU. Monitoring and management of antipsychotic-related metabolic and endocrine adverse events in pediatric patients. *Int Rev Psychiatry* 2008;20:195-201.
67. Goeb JL, Marco S, Duhamel A et al. Metabolic side effects of risperidone in children and adolescents with early-onset schizophrenia. *Prim Care Companion J Clin Psychiatry* 2008;10:486-487.
68. Pringsheim T, Panagiotopoulos C, Davidson J, Ho J. Evidence-based recommendations for monitoring safety of second generation antipsychotics in children and youth. *J Can Acad Child Adolesc Psychiatry* 2011;20:218-233.
69. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J Clin Epidemiol* 2011;64:749-759.
70. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons; 1987.

71. The MI Procedure. *SAS/STAT® 9.3 User's Guide*. SAS Institute Inc; 2011.
72. Schenker N, Taylor S. Partially parametric techniques for multiple imputation. *Computational statistics and data analysis*. 1996. 425-446.
73. Harrell FE. *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis*. New York: Springer; 2015.
74. Harrell FE. Hmisc: A package of miscellaneous R functions. http://biostat.mc.vanderbilt.edu/Hmisc. 2015. Ref Type: Online Source
75. Ambrosioni E, Borghi C, Magnani B. The effect of the angio-tensin-converting-enzyme inhibitor zofenopril on mortality and morbidity after anterior myocardial infarction. *N Engl J Med* 1995;332:80-85.
76. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: a randomizsed factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. *Lancet* 1995;345:669-685.
77. Lewis EJ, Hunsicker LG, Clarke WR et al. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N Engl J Med* 2001;345:851-860.
78. Gruppo Italiano per lo Studio della Sopravvivenza nell'infart Miocardico. GISSI-3: effects of lisinopril and transdermal glyceryl trinitrate singly and together on 6-week mortality and ventricular function after myocardial infarction. *Lancet* 1994;343:1115-1122.
79. Chobanian AV, Bakris GI, Black HR, Cushman WC, Green LA, Izzo JL. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *JAMA* 2003;39(suppl 1):S94-S98.
80. Ahuja TS, Freeman DJr, Mahnken JD, Agraharkar M, Siddiqui M, Memon A. Predictors of the development of hyperkalemia in patients using angiotensin-converting enzyme inhibitors. *Am J Nephrol* 2000;20(4):268-272.
81. Raebel MA, McClure DL, Simon SR et al. Laboratory monitoring of angiotensin converting enzyme inhibitors and angiotensin receptor blockers in ambulatory patients. *Pharmacoepidemiol Drug Saf* 2007;16:55-64.
82. Raebel MA, Ross C, Xu S et al. Diabetes and Drug-Associated Hyperkalemia: Effect of Serum Potassium Monitoring. *J Gen Intern Med* 2010;25(4):326-333 DOI: 10.1007/s11606-009-1228-x.
83. Raebel MA. Hyperkalemia associated with use of angiotensin converting enzyme inhibitors and angiotensin receptor blockers. *Cardiovasc Ther* 2012;30(3):e156-e166.
84. Palmer BF. Managing hyperkalemia caused by inhibitors of the renin-angiotensin-aldosterone system. *N Engl J Med* 2004;351:585-592.
85. Sterne JAC, White IA, Carlin JB et al. Multiple imputation for missing data in epidemiological and clinical research: potentials and pitfalls. *BMJ* 2009;338:b2393.
86. Roy J, Hennessy S. Bayesian hierarchical pattern mixture models for comparative effectiveness of drugs and drug classes using healthcare data: a case study involving antihypertensive medications. *Stats Biosciences* 2011;3(1):79-93.
87. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health* 2010;13(2):273-277. PMID: 19912596.
88. Tsai TT, Ho M, Xu S, et al. Increased risk of bleeding in patients on clopidogrel therapy after drug-eluting stents implantation: insights from the HMO research network-stent registry (HMORN-Stent). *Circ Cardiovasc Interv* 2010;3:230-235.

89. Witt DM, Delate T, Clark NP, et al. Nonadherence with INR monitoring and anticoagulant complications. *Thrombosis Res* 2013;e124-e130.
90. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 2011;10(2):150-161.
91. Toh S, Shetterly S, Powers JD, Arterburn D. Privacy-preserving analytic methods for multisite comparative effectiveness and patient-centered outcomes research. *Med Care* 2014;52(7):664-668.
92. American Pregnancy Association. Human chorionic gonadotropin (HCG); the pregnancy hormone. Available at:http://americanpregnancy.org/while-pregnant/hcg-levels/. 7-2-0015. Ref Type: Online Source
93. Raebel MA, Ellis JL, Andrade SE. Evaluation of gestational age and admission date assumptions used to determine prenatal drug exposure from administrative data. *Pharmacoepidemiol Drug Saf* 2005;14:829-836.
94. Johnsen SL, Wilsgaard T, Rasmussen S, Hanson MA, Godfrey KM, Kiserud T. Fetal size in the second trimester is associated with the duration of pregnancy, small fetuses having longer pregnancies. *BMC Pregnancy Childbirth* 2008;8:25.
95. Margulis AV, Setoguchi S, Mittleman MA, Glynn RJ, Dormuth CR, Hernandez-Diaz S. Algorithms to estimate the beginning of pregnancy in administrative databases. *Pharmacoepidemiol Drug Saf* 2013;22(1):16-24.
96. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009;150:604-612.
97. Rao MK, Morris CD, O'Malley JP, Davis MM, Mori M, Anderson S. Documentation and management of CKD in rural primary care. *CJASN* 2013;8(5):739-748.
98. Castellsague J, Riera-Guaria N, Calingaert B et al. Individual NSAIDs and upper gastrointestinal complications: a systematic review and meta-analysis of observational studies (the SOS Project). *Drug Safety* 2012;35(12):1127-1146.
99. Zhang Z. Interval censoring. *Statistical Methods in Medical Research* 2010;18:53-70.
100. Twisk J, de Boer M, de Vente W, Heymans M. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *J Clin Epidemiol* 2013;66:1022-1028.
101. Allison P. Sensitivity analysis for Not Missing at Random [Web log post]. Retrieved December 21, 2015, from http://statisticalhorizons.com/2014/09. 2014. Ref Type: Online Source
102. Yuan Y. Sensitivity analysis in multiple imputation for missing data. Retrieved December 21, 2015, from http://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf. 2014. SAS Institute Inc, paper SAS270-2014. Ref Type: Online Source

# X. APPENDICES

A. Appendix A. Literature Review Search Terms and Numbers of Articles Identified in Each Search Approach
B. Appendix B. Screenshot of Relational Database Entry Form
C. Appendix C. Copy of Project-Specific Contents of Relational Database
D. Appendix D. Specification Documents for Test Cases
E. Appendix E. Baseline Confounder Test Case 3, Baseline INR in Warfarin Users Starting an Antimicrobial Medication, Standardized Differences Supplementary Tables
F. Appendix F. Additional Results from 5 and from 20 Imputed Datasets for Baseline Confounder Test Case 2
G. Appendix G. Pregnancy-related Diagnosis and Procedure Codes
H. Appendix H. 2012 CKD Cohort Identification and Cohort Augmentation Tables by Data Partner Site
I. Appendix I.

*Appendices can be found in an accompanying document.